A new V-fold type procedure based on robust tests

Lucien Birgé, Nelo Magalhães, Pascal Massart

A new V-fold type procedure based on robust tests

Lucien Birgé*[1], Nelo Magalhães†[2,3], and Pascal Massart‡[2,3]

[1]LPMA, UPMC Université Paris 06
[2]Équipe Probabilités et Statistiques, Université Paris-Sud 11
[3]INRIA team Select

June 2015

Abstract

We define a general V-fold cross-validation type method based on robust tests, which is an extension of the hold-out defined by Birgé [7, Section 9]. We give some theoretical results showing that, under some weak assumptions on the considered statistical procedures, our selected estimator satisfies an oracle type inequality. We also introduce a fast algorithm that implements our method. Moreover we show in our simulations that this V-fold performs generally well for estimating a density for different sample sizes, and can handle well-known problems, such as binwidth selection for histograms or bandwidth selection for kernels. We finally provide a comparison with other classical V-fold methods and study empirically the influence of the value of V on the risk.

1 Introduction

The purpose of this paper is to offer a new method to solve the following problem. Suppose we are given i.i.d. observations from an unknown distribution P_s to be estimated. This distribution is often assumed to have a density s with respect to some given measure μ, hence our notation, but we shall also consider the case when P_s is not absolutely continuous with respect to μ, keeping the same notation P_s for the true distribution, in which case the subscript s just indicates that P_s is the distribution of the observations.

We also have at hand a family of statistical procedures or algorithms $\{\mathcal{A}_m, m \in \mathcal{M}\}$ that can be applied to the observations in order to derive estimators of P_s. How can we use our data in order to choose one potentially optimal algorithm in the family, provided that a criterion of quality for the estimators has been chosen? Let us now be somewhat more precise.

*lucien.birge@upmc.fr
†nelo.moltermagalhaes@gmail.com
‡pascal.massart@math.u-psud.fr

0 *Key words and phrases.* T-estimation, density estimation, V-fold cross-validation, Hellinger loss.

1.1 The problem of procedure choice

We observe an n-sample $\mathbf{X} = \{X_1, \ldots, X_n\}$ of random variables X_i with values in the measured space $(\mathcal{X}, \mathcal{E})$ and we assume (temporarily) that the distribution $P_s = s \cdot \mu$ of the X_i admits a density s with respect to some given positive measure μ on $\mathcal{X}$ and that s belongs to some given subset $\mathcal{S}$ of $\mathbb{L}_1(\mu)$. The purpose here is to use the observations in order to design an estimator $\widehat{s} = \widehat{s}(\mathbf{X})$ of s.

There is a huge amount of strategies for solving this estimation problem, depending on the additional assumptions one makes about s. We shall use the notion of *statistical procedure* (procedure for short), also denoted *statistical algorithm* in what follows, in order to properly formalize these strategies. Following [1], we define a *procedure* or an *algorithm* as any measurable mapping $\mathcal{A}$ from $\bigcup_{k\geq 1} \mathcal{X}^k$ to $\mathcal{S}$. Such a procedure associates to any random sample $\mathbf{Y}_k \in \mathcal{X}^k$ an estimator $\widehat{s}_k = \mathcal{A}(\mathbf{Y}_k) \in \mathbb{L}_1(\mu)$ of s. A classical criterion from decision theory used to measure the quality of a procedure $\mathcal{A}$ based on an i.i.d. sample of size k when s obtains is its *risk*: $\mathbb{E}_s\left[\ell(s, \mathcal{A}(\mathbf{Y}_k))\right]$, where ℓ is some given *loss function* and $\mathbb{E}_s$ denotes the expectation when s obtains, i.e. when the distribution of $\mathbf{Y}_k$ is $P_s^{\otimes k}$. The smaller the risk, the better the procedure $\mathcal{A}$.

To define the risk of a procedure one can consider various loss functions. Some popular ones are derived from a *contrast function* γ (see [9, Definition 1]) which is a mapping from $\mathcal{S} \times \mathcal{X}$ to $\mathbb{R}$ such that s minimizes over $\mathcal{S}$ the function $t \mapsto \mathbb{E}_s\left[\gamma(t, X)\right]$. The loss ℓ at t is then defined as

$$\ell(s, t) = \mathbb{E}_s\left[\gamma(t, X) - \gamma(s, X)\right] \geq 0 \quad \text{for all } t \in \mathcal{S}, \tag{1}$$

hence $\ell(s, s) = 0$. The $\mathbb{L}_2$-loss derives from the choice $\mathcal{S} = \mathbb{L}_2(\mu) \cap \mathbb{L}_1(\mu)$ and $\gamma(t, x) = \|t\|^2 - 2t(x)$, where $\|t\| = \left[\int_{\mathcal{X}} t^2 d\mu\right]^{1/2}$ denotes the $\mathbb{L}_2$-norm. The Kullback-Leibler loss corresponds to the contrast function $\gamma(t, x) = -\log(t(x))$ with $\mathcal{S}$ being the set of all probability densities with respect to μ.

In this paper, we consider the problem of *procedure selection*. Let $(\mathcal{A}_m)_{m\in\mathcal{M}}$ denote a collection of candidate statistical procedures. Our goal is to choose from the observations $\mathbf{X}$ one of these procedures, that is some $\widehat{m}(\mathbf{X}) \in \mathcal{M}$, in order to have the most accurate estimation of s. If we apply all these procedures to the sample $\mathbf{X}$ we get the corresponding collection of estimators $\{\widehat{s}_m = \mathcal{A}_m(\mathbf{X}),\, m \in \mathcal{M}\}$. Given a loss ℓ, the best possible choice for m would be to select $m^* \in \mathcal{M}$ such that

$$\mathbb{E}_s\left[\ell\left(s, \widehat{s}_{m^*}(\mathbf{X})\right)\right] = \inf_{m\in\mathcal{M}} \mathbb{E}_s\left[\ell\left(s, \widehat{s}_m(\mathbf{X})\right)\right].$$

Unfortunately, since s is unknown, all the risks $\mathbb{E}_s\left[\ell(s, \widehat{s}_m)\right]$ are unknown as well and we cannot select the so-called *oracle algorithm* $\mathcal{A}_{m^*}$. One can only hope to choose $\widehat{m} = \widehat{m}(\mathbf{X})$ in such a way that $\mathbb{E}_s\left[\ell(s, \widehat{s}_{\widehat{m}})\right]$ is close to $\mathbb{E}_s\left[\ell(s, \widehat{s}_{m^*})\right]$.

To make this presentation more explicit, let us mention some classical estimation problems that naturally fit into it:

- *Bandwidth selection* (see [11, Chapter 11]). Let $\mathcal{X} = \mathbb{R}$, μ be the Lebesgue measure, $K : \mathbb{R} \to \mathbb{R}$ a given nonnegative function satisfying $\int_{\mathcal{X}} K(x)\, dx = 1$ and $\mathcal{H} = \{h_m, m \in \mathcal{M}\}$ be a finite or countable set of positive bandwidths. We define the *kernel algorithm* $\mathcal{A}_m$ as the procedure that produces from any sample $\mathbf{Y}_k$ of size k a kernel density estimator

with bandwidth h_m, which means that

$$\mathcal{A}_m(\mathbf{Y}_k)(x) = \frac{1}{kh_m} \sum_{Y_i \in \mathbf{Y}_k} K\left(\frac{x - Y_i}{h_m}\right) \quad \text{for all } x \in \mathbb{R}.$$

The problem of choosing a best estimator among the family $\{\widehat{s}_m,\, m \in \mathcal{M}\}$ amounts to select a "best" bandwidth in $\mathcal{H}$, that is one that minimizes the risk $\mathbb{E}_s\left[\ell(s, \widehat{s}_m)\right]$ with respect to m.

- *Model selection* (see [16]). We recall that a *model* S for s is any subset of $\mathcal{S}$. It follows from (1) that minimizing, for t in S, the loss $\ell(s,t)$ derived from the contrast function γ amounts to minimizing $t \mapsto \mathbb{E}_s\left[\gamma(t, X)\right]$ over S. Since s is unknown, this is impossible but if we replace $\mathbb{E}_s\left[\gamma(t, X)\right]$ by its unbiased empirical version: $\gamma_n(t) = n^{-1}\sum_{i=1}^n \gamma(t, X_i)$ we can derive an estimator with values in S by minimizing $\gamma_n(t)$ over S instead. This procedure $\mathcal{A}_S$ is a *minimum contrast algorithm* that provides a *minimum contrast estimator* $\widehat{s}_S(\mathbf{X}) \in \operatorname{argmin}_{t\in S} \gamma_n(t)$ on S. Using for instance, the Kullback-Leibler contrast on a set S of densities leads to the so-called "maximum likelihood estimator" on S.

  If we have at hand some finite or countable collection of models $\{S_m\}_{m\in\mathcal{M}}$ and a suitable contrast function γ we may associate in this way to each model S_m a minimum contrast algorithm $\mathcal{A}_m$ and the corresponding minimum contrast estimator $\widehat{s}_m(\mathbf{X})$. The problem of "model selection" is to select from the data a "best model" (one with the minimal risk) in the family, leading to a "best" possible minimum contrast estimator.

Instead of deriving the loss function ℓ from a contrast function we may use for ℓ the squared Hellinger distance provided that our estimators $\widehat{s}_m$ are genuine probability densities. We recall that the *Hellinger distance* h and the *Hellinger affinity* ρ between two probabilities P and Q defined on $\mathcal{X}$ are given respectively by

$$h(P,Q) = \left[\frac{1}{2}\int\left(\sqrt{dP} - \sqrt{dQ}\right)^2\right]^{1/2} \quad \text{and} \quad \rho(P,Q) = \int\sqrt{dP\,dQ} = 1 - h^2(P,Q), \tag{2}$$

where dP and dQ denote the densities of P and Q with respect to any dominating measure (the result being independent of this choice). One advantage of this loss function lies in the fact that h is a distance on the set $\mathcal{P}$ of *all* probabilities on $\mathcal{X}$ and therefore does not require that P_s be absolutely continuous with respect to μ, which is one of the reasons why we shall use it in the sequel. In this case we take for $\mathcal{S}$ a set of probability densities with respect to μ and we set, for all t in $\mathcal{S}$ and $P_t = t\cdot\mu$, $\ell(s,t) = h^2(P_s, P_t)$ which we shall write $h^2(s,t)$ for simplicity. We shall also write $\rho(t,u)$ for $\rho(P_t, P_u)$. This loss then leads to the *quadratic Hellinger risk*.

1.2 Cross-validation

The biggest difficulty for selecting a procedure in a given family $\{\mathcal{A}_m,\, m \in \mathcal{M}\}$ comes from the fact that we use the same data $\mathbf{X}$ to build the estimators $\widehat{s}_m(\mathbf{X})$ and to evaluate their quality. It is indeed well-known that evaluating the statistical performance of a procedure with the same data that have been used for the construction of the corresponding estimator leads to an overoptimistic result. One solution to avoid this drawback is to save a fraction of

the initial sample to test the output of the procedures $\mathcal{A}_m$ on it. This is the basic idea behind *cross-validation* (CV) which relies on data splitting.

The simplest CV method is the *hold-out* (HO) which corresponds to a single split of the data. The set $\mathbf{X}$ is divided once and for all into two non-empty proper subsets $\mathbf{X}^t$ and $\mathbf{X}^v = \mathbf{X} \setminus \mathbf{X}^t$ to be called respectively the *training* and the *validation* sample. First, with the training sample $\mathbf{X}^t$, we construct a set $\{\mathcal{A}_m(\mathbf{X}^t),\, m \in \mathcal{M}\}$ of preliminary estimators. Then, using the validation sample $\mathbf{X}^v$, we choose a criterion in order to evaluate the quality of each procedure $\mathcal{A}_m$ from the observation of $\mathcal{A}_m(\mathbf{X}^t)$. Finally, we select $\widehat{m}(\mathbf{X}^v)$ minimizing this criterion over $\mathcal{M}$. Depending on the author, the final estimator might be either $\mathcal{A}_{\widehat{m}}(\mathbf{X}^t)$ (as in [11]) or $\mathcal{A}_{\widehat{m}}(\mathbf{X})$ (as in [2]). All CV methods are deduced from the HO: instead of using one single partition of our sample, we use different partitions, compute the HO criterion for each one and finally define the CV criterion by averaging all the HO criteria. The goal, by considering several partitions instead of one, is to reduce the variability with the hope that the CV criterion will lead to a more accurate evaluation of the quality of each procedure.

We shall focus here on V-fold cross-validation (VFCV) which corresponds to a particular set of data splits[1]. One divides the sample $\mathbf{X}$ into $V \geq 2$ disjointed and therefore independent subsamples $\mathbf{X}_j$, $j = 1, \ldots, V$, of the same size $p = n/V$ (assuming, for simplicity, that p is an integer) so that $\mathbf{X} = \bigcup_{j=1}^{V} \mathbf{X}_j$. For each split $j \in \{1, \ldots, V\}$, one uses $\mathbf{X}_j^c$ to build the family of "partial estimators" $\{\widehat{s}_{m,j} = \mathcal{A}_m(\mathbf{X}_j^c),\, m \in \mathcal{M}\}$ and the corresponding validation sample $\mathbf{X}_j$ to define an evaluation criterion $\mathrm{crit}_j(m) = \mathrm{crit}_j(m)(\mathbf{X}_j)$ of the procedure $\mathcal{A}_m(\mathbf{X}_j^c)$ corresponding to the partition $(\mathbf{X}_j, \mathbf{X}_j^c)$ of the data. One finally selects a strategy $\widehat{m}_{\mathrm{VF}}$ minimizing the averaged criterion:

$$\widehat{m}_{\mathrm{VF}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \mathrm{crit}(m) \qquad \text{with} \qquad \mathrm{crit}(m) = \frac{1}{V} \sum_{j=1}^{V} \mathrm{crit}_j(m).$$

There are as many V-fold procedures as there are different ways to define $\mathrm{crit}_j(m)$. If we work with a loss of the type (1), the best estimator in the family $\{\widehat{s}_{m,j},\, m \in \mathcal{M}\}$ is the one minimizing the loss, i.e. the one minimizing $\mathbb{E}_s\left[\gamma(\widehat{s}_{m,j}, X)\right]$ (with X being independent of $\mathbf{X}_j^c$). A natural idea for evaluating this quantity that we cannot compute since we do not know s is to estimate it by its empirical version based on the independent sample $\mathbf{X}_j$ of size p, which leads to the criterion

$$\mathrm{crit}_j(m) = \frac{1}{p} \sum_{X_i \in \mathbf{X}_j} \gamma\left(\widehat{s}_{m,j}, X_i\right).$$

In this classical context, we naturally select the statistical procedure with the lowest estimated average loss $\mathrm{crit}(m)$. The choice $\gamma(t,x) = -\log(t(x))$ leads to the Kullback-Leibler V-fold (KLVF) whereas $\gamma(t,x) = \|t\|^2 - 2t(x)$ provides the Least-Squares V-fold (LSVF). The chosen estimators will be respectively denoted $\widehat{m}_{\mathrm{KLVF}}$ and $\widehat{m}_{\mathrm{LSVF}}$ and the relevant classical criterion will be denoted $\mathrm{crit}_{\mathrm{VFCV}}$ in what follows.

1.3 An alternative criterion

When the chosen loss function that we use is the squared Hellinger distance, an alternative empirical criterion to evaluate the quality of an estimator has been proposed by Birgé [5]

[1] The concerned reader should have a look at the survey of Arlot and Celisse [1] to get a complete overview of other CV methods.

following ideas of Le Cam [12, 13] to process estimator selection. An alternative method was later introduced by Baraud [3]. An HO strategy based on this criterion was first proposed by Birgé in [7], this latter procedure being recently implemented in [14]. The idea behind the construction is as follows. Suppose we have at hand a set $\mathcal{T}$ of densities with respect to μ and, for each pair (t,u), $t \neq u$, of points of $\mathcal{T}$, a test $\psi_{t,u}$ between t and u ($\psi_{t,u} = t$ meaning accepting t). Given a sample $\mathbf{X}$ we may perform all the tests $\psi_{t,u}(\mathbf{X})$ and consider the criterion $\mathcal{D}(t)$ defined on $\mathcal{T}$ by

$$\mathcal{D}(t) = \sup_{u \in \mathcal{T},\, u \neq t} h(t,u) \mathbb{1}_{\{\psi_{t,u}(\mathbf{X}) = u\}}. \tag{3}$$

It immediately follows from this definition that

$$h(t,u) \leq \max\{\mathcal{D}(t), \mathcal{D}(u)\} \quad \text{for all } t,u \in \mathcal{T}. \tag{4}$$

This definition means that $\mathcal{D}(t)$ is large when there exists some u which is far from t and which is preferred to t by the test $\psi_{t,u}(\mathbf{X})$, suggesting that t is likely to be far from s, at least if s does belong to $\mathcal{T}$. In order that this be actually true even if P_s does not belong to $\{P_t, t \in \mathcal{T}\}$, it is necessary to design suitable tests. It has been shown in [5] that one can build a special test $\psi_{t,u}$ between the two Hellinger balls $\mathcal{B}(t,r)$ and $\mathcal{B}(u,r)$ with $r < h(t,u)/2$ (where $\mathcal{B}(t,r)$ denotes the closed ball of center t and radius r in the metric space $(\mathcal{P}, h)$) which posesses the required properties. With this special choice of tests $\psi_{t,u}$ for all pairs (t,u), $\mathcal{D}(t)$ becomes indeed a good indicator of the quality of t as an estimator of s (the smaller $\mathcal{D}(t)$, the better t) and, more generally, of P_t as an estimator of P_s even if P_s is not absolutely continuous with respect to μ. This property of $\mathcal{D}$ suggests to define the following criterion on which to base a new VFCV procedure. Starting from the family of preliminary density estimators

$$\left\{ \widehat{s}_{m,j} = \mathcal{A}_m(\mathbf{X}_j^c),\ m \in \mathcal{M},\ 1 \leq j \leq V \right\},$$

we build all the corresponding tests $\psi_{\widehat{s}_{l,j}, \widehat{s}_{m,j}}(\mathbf{X}_j)$, hereafter denoted for simplicity by $\psi_{l,m}(\mathbf{X}_j)$, between the densities $\widehat{s}_{l,j}$ and $\widehat{s}_{m,j}$ for $l, m \in \mathcal{M}$, $l \neq m$. Then, for each j and m, we define the criterion $\mathrm{crit}_j(m)$ by

$$\mathrm{crit}_j(m) = \mathcal{D}_j^2(m) \quad \text{with} \quad \mathcal{D}_j(m) = \sup_{l \in \mathcal{M},\, l \neq m} h(\widehat{s}_{l,j}, \widehat{s}_{m,j}) \mathbb{1}_{\{\psi_{l,m}(\mathbf{X}_j) = l\}}. \tag{5}$$

We then naturally define our test-based V-fold criterion as

$$\mathrm{crit}_{\mathrm{TVF}}(m) := \overline{\mathcal{D}}^2(m) = \frac{1}{V} \sum_{j=1}^{V} \mathcal{D}_j^2(m) \quad \text{for all } m \in \mathcal{M}.$$

Up to our knowledge, this is the first V-fold type procedure based on the Hellinger distance. Note that this construction requires that the estimators $\widehat{s}_{m,j}$ be genuine probability densities with respect to μ which we shall assume from now on.

1.4 Organization of the paper

Our goal is to study our new VFCV procedure from both a theoretical and a practical point of view. Section 2 is dedicated to its theoretical study. In Section 3 we discuss in details the implications of the resulting risk bounds to the case of histogram estimators, applications

to kernel estimators and an extension to algorithms that do not lead to genuine probability density estimators. Section 4 contains an empirical study of the influence of the value of V on the performance of our procedure in terms of Hellinger risk and also comparisons with classical V-fold and some especially calibrated procedures. Section 5 describes the fast algorithm that we have designed and implemented in order to compute the selected estimator efficiently. Finally Section 6 contains a proof of the bounds for the Hellinger risk of kernel estimators. We provide some additional simulations in Section A.

2 T-V-fold

As already mentioned, the method proposed in [7] is based on tests and it results in what Birgé called T-estimators (T for "test"). We shall therefore call our cross-validation method based on the same tests T-V-fold cross-validation (TVF for short).

2.1 Tests between Hellinger balls

The tests that we use for our procedure satisfy the following assumption, which ensures their robustness. We recall that $\mathcal{S}$ is the set of all probability densities with respect to μ.

Assumption (TEST). *Let $\theta \in (0, 1/2)$ be given. For all t and u in $\mathcal{S}$, $z \in \mathbb{R}$ and $r = \theta h(t,u)$ there exists some test statistic $T_{t,u,\theta}(\mathbf{X})$ depending on t, u, θ and $\mathbf{X}$ with the following properties. The test $\psi_{t,u}$ between t and u defined by*

$$\psi_{t,u}(\mathbf{X}) = \begin{cases} t & \text{if} \quad T_{t,u,\theta}(\mathbf{X}) > z \\ u & \text{if} \quad T_{t,u,\theta}(\mathbf{X}) < z \end{cases}, \quad z \in \mathbb{R}, \tag{6}$$

with an arbitrary choice when $T_{t,u,\theta}(\mathbf{X}) = z$, satisfies

$$\sup_{\{\, P_s \in \mathcal{P} \,|\, h(s,t) \le r \,\}} \mathbb{P}_s \left[\psi_{t,u}(\mathbf{X}) = u \right] \le \exp\left[-n(1-2\theta)^2 h^2(t,u) + z \right] \tag{7}$$

and

$$\sup_{\{\, P_s \in \mathcal{P} \,|\, h(s,u) \le r \,\}} \mathbb{P}_s \left[\psi_{t,u}(\mathbf{X}) = t \right] \le \exp\left[-n(1-2\theta)^2 h^2(t,u) - z \right], \tag{8}$$

where $\mathbb{P}_s$ denotes the probability that gives $\mathbf{X}$ the distribution $P_s^{\otimes n}$.

Any test satisfying (7) and (8) will be suitable for our needs.

Tests between balls In order to define tests between two Hellinger balls $\mathcal{B}(t,r)$ and $\mathcal{B}(u,r)$ with $r = \theta h(t,u)$, $0 < \theta < 1/2$, Birgé introduced the following test statistic

$$T_{t,u,\theta}(\mathbf{X}) = \sum_{i=1}^{n} \log\left(\frac{\sin(\omega(1-\theta))\sqrt{t}(X_i) + \sin(\omega\theta)\sqrt{u}(X_i)}{\sin(\omega(1-\theta))\sqrt{u}(X_i) + \sin(\omega\theta)\sqrt{t}(X_i)} \right) \quad \text{with } \omega = \arccos \rho(t,u). \tag{9}$$

We should notice that for $\theta = 0$, the test given by (9) is exactly the likelihood ratio test between t and u. The fact that Assumption (TEST) holds for this test whatever $\theta \in (0, 1/2)$ has been proven in [6] and a more up-to-date version is to be found in [8, Corollary 1].

2.2 TVF estimators

Let $(\Delta_m)_{m\in\mathcal{M}}$ denote some collection of positive numbers satisfying

$$\Delta_m \geq 0 \quad \text{for all } m \in \mathcal{M}, \qquad \text{and} \qquad \frac{1}{2} \leq \Gamma = \sum_{m\in\mathcal{M}} \exp(-\Delta_m) < \infty. \tag{10}$$

Starting from the family of estimators $\widehat{s}_{m,j}$ defined in Section 1.3, we consider the corresponding tests $\psi_{l,m}(\mathbf{X}_j) = \psi_{\widehat{s}_{l,j},\widehat{s}_{m,j}}(\mathbf{X}_j)$ with $t = \widehat{s}_{l,j}$, $u = \widehat{s}_{m,j}$ and $z = \Delta_l - \Delta_m$ in (6). This results in the estimator $\widehat{s}_{\widehat{m}_{\mathrm{TVF}}}$ derived from the procedure $\mathcal{A}_{\widehat{m}_{\mathrm{TVF}}}$ with

$$\widehat{m}_{\mathrm{TVF}} \in \operatorname*{argmin}_{m\in\mathcal{M}} \overline{\mathcal{D}}^2(m) = \operatorname*{argmin}_{m\in\mathcal{M}} \frac{1}{V}\sum_{j=1}^{V} \mathcal{D}_j^2(m). \tag{11}$$

2.3 Assumption on the family of procedures

The idea of V-fold relies on the heuristic that, for each procedure $\mathcal{A}_m$, the observation of V partial estimators $\widehat{s}_{m,j}$, $1 \leq j \leq V$ based on samples of size $n-p$ with $p = n/V$ allows to predict the behavior of an estimator $\widehat{s}_m$ based on an n-sample. This requires that there exists a link between the loss of $\widehat{s}_m$ and the losses of the $\widehat{s}_{m,j}$. We shall need the following assumption on the collection of procedures we consider.

Assumption (LOSS). *For all procedures $\mathcal{A}_m$ with $m \in \mathcal{M}$, the loss at s satisfies*

$$h^2\left(s, \widehat{s}_m\right) \leq \frac{1}{V}\sum_{j=1}^{V} h^2\left(s, \widehat{s}_{m,j}\right).$$

This implies in particular that $R(\mathcal{A}_m, n, s) \leq R(\mathcal{A}_m, n-p, s)$, where

$$R(\mathcal{A}, k, s) = \mathbb{E}_s\left[h^2\left(s, \mathcal{A}(\mathbf{Y}_k)\right)\right]$$

denotes the risk at s of the procedure $\mathcal{A}$ based on a sample of size k. Assumption (LOSS) is in particular satisfied by the "additive estimators" of [11, Chapter 10].

Definition 1. An *additive estimator* $\widehat{s} = \widehat{s}(\mathbf{X})$ derived from a sample $\mathbf{X}$ of size n is an estimator that can be written in the form:

$$\widehat{s}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}(x, X_i) \quad \text{for all } x \in \mathcal{X}, \tag{12}$$

where $\mathcal{K}$ is a real valued function from $\mathcal{X}\times\mathcal{X}$ to $\mathbb{R}$.

There is a huge amount of literature about these estimators which already appeared in an early version in Whittle [21]. The first results about their asymptotic properties in general were made by Watson and Leadbetter [20], followed by Winter [22] and Walter and Blum [19] who established rates (the latter authors called them *delta sequence density estimators*). They were introduced in the context of CV by Rudemo [18] and used by Marron for comparison of CV techniques [15]. As shown in [19] and [11], additive estimators include in particular:

- *Histogram estimators.* Given a partition $\{I_\lambda,\, \lambda \in \Lambda\}$ of $\mathcal{X}$ with $0 < \mu(I_\lambda) < +\infty$ for all λ one defines the histogram estimator based on this partition as

$$\widehat{s}(x) = \sum_{\lambda \in \Lambda} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{I_\lambda}(X_i) \right) \frac{\mathbb{1}_{I_\lambda}(x)}{\mu(I_\lambda)}. \tag{13}$$

  It corresponds to the case of $\mathcal{K}(x, X_i) = \sum_{\lambda \in \Lambda} [\mu(I_\lambda)]^{-1} \mathbb{1}_{I_\lambda}(X_i) \mathbb{1}_{I_\lambda}(x)$.

- *Parzen kernel estimators on the line.* Set $\mathcal{K}(x, X_i) = w^{-1} K\left(w^{-1}(X_i - x)\right)$ for a given nonnegative kernel K with $\int_{\mathbb{R}} K(x)\, dx = 1$ and a positive bandwidth w. Then (12) leads to a density estimator with respect to the Lebesgue measure on $\mathbb{R}$.

It is straightforward to check that if the procedure $\mathcal{A}_m$ results in additive estimators, the following relationship which says that the estimator built with the whole sample is exactly the convex combination of the V partial estimators holds:

$$\widehat{s}_m = \frac{1}{V} \sum_{j=1}^{V} \widehat{s}_{m,j}. \tag{14}$$

As a consequence, we get the following elementary property:

Proposition 1. *Any procedure $\mathcal{A}_m$ which results in additive estimators does satisfy Assumption (LOSS).*

Proof. It follows from (14) and the concavity of the square root function that

$$\rho(s, \widehat{s}_m) = \rho\left(s, \frac{1}{V} \sum_{j=1}^{V} \widehat{s}_{m,j} \right) \geq \frac{1}{V} \sum_{j=1}^{V} \rho(s, \widehat{s}_{m,j}),$$

which is exactly Assumption (LOSS) in view of (2). □

2.4 The main result

Assumption (LOSS) ensures that, for the procedures we consider, the loss of some estimator is bounded by the mean of the losses of the partial estimators. This motivates us to work separately on each split $j \in \{1, \ldots, V\}$ and then to deduce a risk bound for the estimator built with the whole sample. It is therefore natural to study for each j the deviations of the random variable $\mathcal{D}_j(\cdot)$. A deviation inequality for $\mathcal{D}$ has been proven in Theorem 9 of [7]. Let us now recall it and provide a short proof for the sake of completeness.

Proposition 2. *Let $(\Delta_m)_{m \in \mathcal{M}}$ be a collection of weights satisfying (10) and*

$$A = \frac{n(1-2\theta)^2}{2V}; \qquad y_{m,j} = \max\left(\frac{h(s, \widehat{s}_{m,j})}{\theta}, \sqrt{\frac{\Delta_m}{A}} \right).$$

Then, for all $m \in \mathcal{M}$, and $j \in \{1, \ldots, V\}$,

$$\mathbb{P}_s\left[\mathcal{D}_j(m) \geq y \,\middle|\, \mathbf{X}_j^c \right] \leq \Gamma \exp\left[-2Ay^2 + \Delta_m \right] \qquad \text{for all } y \geq y_{m,j}.$$

Proof. Let us fix some $m \in \mathcal{M}$ and $j \in \{1, \dots, V\}$ and work conditionally to the training sample $\mathbf{X}_j^c$ so that the collection of estimators $(\widehat{s}_{l,j})_{l\in\mathcal{M}}$ can be considered as fixed. We perform the test $\psi_{l,m}(\mathbf{X}_j)$ that satisfy Assumption (TEST) with $z = \Delta_l - \Delta_m$ in (7). Then

$$\begin{aligned}
\mathbb{P}_s\left[\mathcal{D}_j(m) \geq y \mid \mathbf{X}_j^c\right] &= \mathbb{P}_s\left[\exists\, l \in \mathcal{M} \text{ such that } h(\widehat{s}_{l,j}, \widehat{s}_{m,j}) \geq y \text{ and } \psi_{l,m}(\mathbf{X}_j) = l \mid \mathbf{X}_j^c\right] \\
&\leq \sum_{l\in\mathcal{M}:\ h(\widehat{s}_{l,j},\widehat{s}_{m,j})\geq y} \mathbb{P}_s\left[\psi_{l,m}(\mathbf{X}_j) = l \mid \mathbf{X}_j^c\right] \\
&\leq \sum_{l\in\mathcal{M}:\ h(\widehat{s}_{l,j},\widehat{s}_{m,j})\geq y} \exp\left[-2Ah^2\left(\widehat{s}_{l,j}, \widehat{s}_{m,j}\right) - \left(\Delta_l - \Delta_m\right)\right] \\
&\leq \exp\left[-2Ay^2 + \Delta_m\right] \sum_{l\in\mathcal{M}} \exp(-\Delta_l) \quad \leq \quad \Gamma \exp\left[-2Ay^2 + \Delta_m\right],
\end{aligned}$$

where we successively used the fact that $y \geq y_{m,j} \geq \theta^{-1} h(s, \widehat{s}_{m,j})$ and (10). □

For each fixed j, that is conditionally to each $\mathbf{X}_j^c$, we deal with some "fixed geometrical configuration" since the points $(\widehat{s}_{m,j})_{m\in\mathcal{M}}$ are given, conditionally to $\mathbf{X}_j^c$. On this configuration, Proposition 2 controls the deviations of $\mathcal{D}_j^2(m)$ which allows us to bound the expectation of $\overline{\mathcal{D}}^2(m)$. This results in the following theorem.

Theorem 1. *Under Assumption (LOSS), the estimator* $\widehat{s}_{\widehat{m}_{\mathrm{TVF}}} = \mathcal{A}_{\widehat{m}_{\mathrm{TVF}}}(\mathbf{X})$ *with* $\widehat{m}_{\mathrm{TVF}}$ *minimizing the criterion* $\overline{\mathcal{D}}^2(m)$ *satisfies the following inequality:*

$$\mathbb{E}_s\left[h^2\left(s, \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}\right)\right] \leq \inf_{m\in\mathcal{M}} \left\{ 2\left(\frac{\theta^2+2}{\theta^2}\right) R\left(\mathcal{A}_m, \frac{V-1}{V}n, s\right) + \frac{4V[\Delta_m + \log(2\Gamma) + 1]}{n(1-2\theta)^2} \right\}. \tag{15}$$

Proof. Let m' be any point in $\mathcal{M}$. It follows from (4) that, for all $m \in \mathcal{M}$ and $1 \leq j \leq V$,

$$h\left(s, \widehat{s}_{m',j}\right) \leq h\left(s, \widehat{s}_{m,j}\right) + h\left(\widehat{s}_{m',j}, \widehat{s}_{m,j}\right) \leq h\left(s, \widehat{s}_{m,j}\right) + \max\left(\mathcal{D}_j(m), \mathcal{D}_j(m')\right).$$

Setting $m' = \widehat{m}_{\mathrm{TVF}} = \widehat{m}$ for short, we derive that

$$\begin{aligned}
\frac{1}{V}\sum_{j=1}^V h^2\left(s, \widehat{s}_{\widehat{m},j}\right) &\leq 2\left\{\frac{1}{V}\sum_{j=1}^V h^2\left(s, \widehat{s}_{m,j}\right) + \frac{1}{V}\sum_{j=1}^V \max\left(\mathcal{D}_j^2(m), \mathcal{D}_j^2(\widehat{m})\right)\right\} \\
&\leq 2\left\{\frac{1}{V}\sum_{j=1}^V h^2\left(s, \widehat{s}_{m,j}\right) + \frac{1}{V}\sum_{j=1}^V (\mathcal{D}_j^2(m) + \mathcal{D}_j^2(\widehat{m}))\right\} \\
&\leq \frac{2}{V}\sum_{j=1}^V h^2\left(s, \widehat{s}_{m,j}\right) + 4\overline{\mathcal{D}}^2(m),
\end{aligned}$$

for all $m \in \mathcal{M}$. Using Assumption (LOSS) and taking expectations, we derive that

$$\mathbb{E}_s\left[h^2\left(s, \widehat{s}_{\widehat{m}}\right)\right] \leq \frac{1}{V}\sum_{j=1}^V \mathbb{E}_s\left[h^2\left(s, \widehat{s}_{\widehat{m},j}\right)\right] \leq 2R(\mathcal{A}_m, n-p, s) + 4\mathbb{E}_s\left[\overline{\mathcal{D}}^2(m)\right], \tag{16}$$

since the risk of $\widehat{s}_{m,j}$ is the same for all j and equal to $R(\mathcal{A}_m, n-p, s)$.

Let now m and j be fixed. Integrating the bound for $\mathbb{P}_s\left[\mathcal{D}_j^2(m) \geq y \,\middle|\, \mathbf{X}_j^c\right]$ provided by Proposition 2 with respect to y leads to

$$\mathbb{E}_s\left[\mathcal{D}_j^2(m) \,\middle|\, \mathbf{X}_j^c\right] \leq y_{m,j}^2 + \Gamma e^{\Delta_m} \int_{y_{m,j}^2}^1 e^{-2Az}\, dz \leq y_{m,j}^2 + \frac{\Gamma e^{\Delta_m}}{A} \exp\left(-2Ay_{m,j}^2\right)$$

and, since $Ay_{m,j}^2 \geq \Delta_m$,

$$\mathbb{E}_s\left[\mathcal{D}_j^2(m)\right] \leq \mathbb{E}_s\left[y_{m,j}^2\right] + \Gamma A^{-1} \exp(-\Delta_m) \leq \frac{1}{\theta^2}\mathbb{E}_s\left[h^2\left(s, \widehat{s}_{m,j}\right)\right] + \frac{\Delta_m + \Gamma e^{-\Delta_m}}{A}.$$

Finally

$$\mathbb{E}_s\left[\overline{\mathcal{D}}^2(m)\right] \leq \frac{1}{\theta^2} R(\mathcal{A}_m, n-p, s) + \frac{\Delta_m + \Gamma e^{-\Delta_m}}{A}.$$

One should then observe that changing Δ_m into $\Delta_m + B$ with $B \geq 0$ does not change the procedure since the tests only depend on differences $\Delta_m - \Delta_l$. Since the new weights $\Delta_m + B$ also satisfy (10) with Γ changed to Γe^{-B}, the previous bound remains valid for the new weights leading to

$$\mathbb{E}_s\left[\overline{\mathcal{D}}^2(m)\right] \leq \frac{1}{\theta^2} R(\mathcal{A}_m, n-p, s) + \frac{\Delta_m + B + \Gamma e^{-\Delta_m - 2B}}{A}.$$

An optimization with respect to B (taking into account the fact that $\Gamma \geq 1/2$) together with (16) leads to our conclusion. □

2.5 Comments

At this stage, several comments are in order:

A simple case It is often the case that $\mathcal{M}$ is finite with cardinality $|\mathcal{M}|$ and that we use equal weights $\Delta_m = \Delta \leq \log(2|\mathcal{M}|)$ for all $m \in \mathcal{M}$, in which case $\Gamma = |\mathcal{M}|e^{-\Delta}$ which leads to the following risk bound which only depends on $|\mathcal{M}|$:

$$\mathbb{E}_s\left[h^2\left(s, \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}\right)\right] \leq 2\left(\frac{\theta^2+2}{\theta^2}\right) \inf_{m \in \mathcal{M}} R\left(\mathcal{A}_m, \frac{V-1}{V}n, s\right) + \frac{4V\log(2e|\mathcal{M}|)}{n(1-2\theta)^2}.$$

Modified V-fold Unfortunately, there are actually many estimators, like maximum likelihood estimators or T-estimators, that do not satisfy Assumption (LOSS) and for which the previous risk computations fail. In order to solve this problem, one should think about the initial purpose of VF methods and, more generally, of procedure selection. Starting from the family $\{\mathcal{A}_m,\, m \in \mathcal{M}\}$, we want to determine, at least approximately, the best procedure for the problem at hand. But if we design an alternative procedure $\overline{\mathcal{A}}$ not contained in the initial set, but as good as the best one in the set, we may consider that we have achieved our goal.

It should be noted at this stage that Assumption (LOSS) is only used to derive in (16) that

$$\mathbb{E}_s\left[h^2\left(s, \widehat{s}_{\widehat{m}}\right)\right] \leq \frac{1}{V}\sum_{j=1}^V \mathbb{E}_s\left[h^2\left(s, \widehat{s}_{\widehat{m},j}\right)\right],$$

which, in view of Proposition 1, holds as soon as $\widehat{s}_{\widehat{m}} = V^{-1}\sum_{j=1}^{V}\widehat{s}_{\widehat{m},j}$. A natural solution to deal with any family of estimators that do not satisfy Assumption (LOSS) is therefore as follows. Define the partial estimators $\widehat{s}_{m_j}$ and determine $\widehat{m}_{\mathrm{TVF}}$ as before by (11), then define the final TVF-estimator $\widetilde{s}_{\mathrm{TVF}}$ by

$$\widetilde{s}_{\mathrm{TVF}} = V^{-1}\sum_{j=1}^{V}\widehat{s}_{\widehat{m}_{\mathrm{TVF}},j} \tag{17}$$

so that (16) is satisfied and the proof proceeds as before; our modified TVF-estimator $\widetilde{s}_{\mathrm{TVF}}$ satisfies the conclusion of Theorem 1.

Extension It should be noted that the following analogue of (15) holds (with the same proof)

$$\mathbb{E}_s\left[h^2\left(s,\widehat{s}_{\widehat{m}_{\mathrm{TVF}}}\right)\right] \le \inf_{m\in\mathcal{M}}\left\{C_1(\theta,a)R\left(\mathcal{A}_m,\frac{V-1}{V}n,s\right)+C_2(\theta,a)\frac{V(\Delta_m+\log(2\Gamma)+1)}{n}\right\},$$

if we replace Assumption (TEST) by the following

Assumption (TEST'). *Let $\theta\in(0,1/2)$ and $a>0$ be given. For all t and u in $\mathcal{S}$ and $r=\theta h(t,u)$ there exists some test statistic $T_{t,u,\theta}(\mathbf{X})$ depending on t,u,θ and $\mathbf{X}$ with the following properties. The test $\psi_{t,u}$ between t and u defined by*

$$\psi_{t,u}(\mathbf{X}) = \begin{cases} t & \text{if} \quad T_{t,u,\theta}(\mathbf{X}) > z \\ u & \text{if} \quad T_{t,u,\theta}(\mathbf{X}) < z \end{cases}, \qquad z\in\mathbb{R},$$

with an arbitrary choice when $T_{t,u,\theta}(\mathbf{X})=z$, satisfies

$$\sup_{\{\,P_s\in\mathcal{P}\,|\,h(s,t)\le r\,\}}\mathbb{P}_s\left[\psi_{t,u}(\mathbf{X})=u\right]\le\exp\left[-nah^2(t,u)+z\right]$$

and

$$\sup_{\{\,P_s\in\mathcal{P}\,|\,h(s,u)\le r\,\}}\mathbb{P}_s\left[\psi_{t,u}(\mathbf{X})=t\right]\le\exp\left[-nah^2(t,u)-z\right],$$

where $\mathbb{P}_s$ denotes the probability that gives $\mathbf{X}$ the distribution $P_s^{\otimes n}$.

In particular Baraud introduced in [3] and for the same purpose of estimator selection the following statistic that relies on a variational formula for the Hellinger affinity. For $r=(t+u)/2$, let

$$T_{t,u}(\mathbf{X}) = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\sqrt{t}(X_i)-\sqrt{u}(X_i)}{\sqrt{r}(X_i)}+\int\left(\sqrt{t(x)}-\sqrt{u(x)}\right)\sqrt{r(x)}\,d\mu(x)\right). \tag{18}$$

The corresponding test $\psi_{t,u}$ actually satisfies Assumption (TEST') for small enough constants θ and a. This follows from Baraud (2008, unpublished manuscript). Therefore the test $\psi(t,u)$ derived from Baraud's statistic could be used instead of the tests between balls. Some simulations based on this alternative test will be provided in Section A.

3 About the choice of V

Let us now come back to the bound (15). It follows from our empirical study in Section 4.2 below that a good choice of θ is $1/4$. Therefore assuming, to be specific and for simplicity, that $\theta = 1/4$ and that

$$\log(2\Gamma) + 1 \leq 3\Delta_m \quad \text{for all } m \in \mathcal{M}, \tag{19}$$

(15) becomes

$$\mathbb{E}_s\left[h^2\left(s, \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}\right)\right] \leq 66 \inf_{m\in\mathcal{M}}\left\{R\left(\mathcal{A}_m, \frac{V-1}{V}n, s\right) + \frac{V\Delta_m}{n}\right\}. \tag{20}$$

Although this risk bound is certainly far from optimal in view of the large constant 66 and our extended simulations show that the actual risk is indeed substantially smaller, it is nevertheless already enlightening. To see it, let us begin with the simple case of regular histograms.

3.1 Regular histograms

Let us analyze the problem of estimating an unknown density s with respect to the Lebesgue measure on $[0,1]$ from n i.i.d. observations with density s. We consider, for each positive integer m, the histogram estimator $\widehat{s}_m$ based on the partition $\mathcal{I}_m$ of $[0,1]$ into m intervals of equal length m^{-1}. It is known from [10, Theorem 1] that the risk at s of the histogram $\widehat{s}_m$ built from n i.i.d. observations is bounded by

$$\mathbb{E}_s\left[h^2\left(s, \widehat{s}_m\right)\right] \leq h^2\left(s, \overline{s}_m\right) + \frac{m-1}{2n}, \tag{21}$$

where $\overline{s}_m$ is the $\mathbb{L}_2$-projection of s onto the m-dimensional linear space of piecewise constant functions on the partition $\mathcal{I}_m$. It is also shown in this theorem that this bound is asymptotically optimal, up to a factor 4, since the asymptotic risk (when n goes to infinity) is of the form

$$\mathbb{E}_s\left[h^2\left(s, \widehat{s}_m\right)\right] = h^2\left(s, \overline{s}_m\right) + \frac{m-1}{8n}\Big(1 + o(1)\Big). \tag{22}$$

In view of (22), the bound in (21) can be considered as optimal, up to a constant factor and it follows from (21) that

$$R\left(\mathcal{A}_m, \frac{V-1}{V}n, s\right) \leq h^2\left(s, \overline{s}_m\right) + \frac{(m-1)V}{2n(V-1)} = h^2\left(s, \overline{s}_m\right) + \frac{m-1}{2n} + \frac{m-1}{2n(V-1)} \tag{23}$$

and that

$$\inf_{m\in\mathcal{M}} \mathbb{E}_s\left[h^2\left(s, \widehat{s}_m\right)\right] \leq h^2\left(s, \overline{s}_{m^*}\right) + \frac{(m^*-1)}{2n} = \inf_{m\in\mathcal{M}}\left\{h^2\left(s, \overline{s}_m\right) + \frac{(m-1)}{2n}\right\}, \tag{24}$$

where this last bound can be considered as a benchmark for the risk of any selection procedure applied to our family of histograms. Since the Hellinger distance is bounded by 1, it clearly appears that one should restrict to values of m that are not larger than $2n$. We shall therefore now assume that $\mathcal{M} = \{1, 2, \ldots, 2n\}$.

Applying (23) to (20), we get

$$\frac{1}{66}\mathbb{E}_s\left[h^2\left(s,\widehat{s}_{\widehat{m}_{\mathrm{TVF}}}\right)\right] \leq \inf_{m\in\mathcal{M}}\left\{R\left(\mathcal{A}_m,\frac{V-1}{V}n,s\right)+\frac{V\Delta_m}{n}\right\} \tag{25}$$

$$\leq \inf_{m\in\mathcal{M}}\left\{\left(h^2\left(s,\overline{s}_m\right)+\frac{m-1}{2n}\right)+\left(\frac{m-1}{2n(V-1)}+\frac{V\Delta_m}{n}\right)\right\} \tag{26}$$

$$\leq \left[h^2\left(s,\overline{s}_{m^*}\right)+\frac{(m^*-1)}{2n}\right]+\left[\frac{m^*-1}{2n(V-1)}+\frac{V\Delta_{m^*}}{n}\right], \tag{27}$$

with m^* defined by (24). We see from (26) that, up to the multiplicative constant 66, we have to optimize with respect to m a bound for the risk of $\widehat{s}_m$ plus a residual term which depends in a non-monotonous way of V. The bound (27) shows that, up to a constant factor, we actually recover our benchmark (24) plus an error term which writes

$$g(V-1) \quad \text{with} \quad g(x)=\frac{1}{n}\left(\frac{m^*-1}{2x}+x\Delta_{m^*}\right)+\frac{\Delta_{m^*}}{n}.$$

Clearly, $g(x)$ is minimum for $x=x_0=\sqrt{(m^*-1)/(2\Delta_{m^*})}$. It follows that the optimal value of V is two if $m^*-1\leq 2\Delta_{m^*}$. This occurs in particular if $m^*=1$, for instance when P_s is the uniform distribution on $[0,1]$ or close enough to it. It also occurs if $\Delta_m\geq(m-1)/2$ for all $m\geq 2$.

Let us now consider the situation for which $m^*-1>2\Delta_{m^*}$ so that $x_0>1$ and the optimal value of V belongs to (x_0-1,x_0+1). If $(m-1)/\Delta_m$ is an increasing function of m, the optimal value of V will be a non-decreasing function of m^* which, as m^* does, depends on the true unknown value of s, large values of m^* leading to large values for V and vice-versa. For instance, the choice of equal weights, $\Delta_m=\log 2n$ for $m\in\mathcal{M}$ leads to $\Gamma=1$ which satisfies (19) and to an optimal V of order $\sqrt{(m^*-1)/(2\log 2n)}$. But this choice of Δ_m is certainly not optimal in view of (25). A better one would be $\Delta_m=(1/3)+2\log m$ which also satisfies (19) but improves (25) substantially. Then the optimal value of V is of order $\sqrt{(m^*-1)/((2/3)+4\log m^*)}$, still depending on the true unknown s. Only larger values of Δ_m of the form $\Delta_m=a(m-1)$ for $m\geq 2$ that deteriorate the bound (25) and therefore should not be recommended lead to an optimal value of V which is independent of m^*, hence of s.

3.2 The typical situation

A risk bound of the form (21) is actually not specific of histograms but actually rather typical. There are many procedures for which the risk, for a convenient choice of the index m and of the set $\mathcal{M}\subset\mathbb{R}$ can be bounded in the following way:

$$\mathbb{E}_s\left[h^2\left(s,\widehat{s}_m\right)\right]\leq H(s,m)+Cmn^{-1}, \tag{28}$$

where H is a nonincreasing function of m, leading to an optimal choice m^* for m (with respect to this bound which we take as a benchmark for the risk) given by

$$m^*=\mathrm{argmin}_{m\in\mathcal{M}}\left\{H(s,m)+Cmn^{-1}\right\}.$$

It then follows from (20) that we get an analogue of (27), namely

$$\begin{aligned}\frac{1}{66}\mathbb{E}_s\left[h^2\left(s,\widehat{s}_{\widehat{m}_{\mathrm{TVF}}}\right)\right] &\leq \inf_{m\in\mathcal{M}}\left\{H(s,m)+\frac{CmV}{n(V-1)}+\frac{V\Delta_m}{n}\right\}\\ &\leq \left[H(s,m^*)+\frac{Cm^*}{n}\right]+\frac{1}{n}\left[\frac{Cm^*}{V-1}+(V-1)\Delta_{m^*}\right]+\frac{\Delta_{m^*}}{n}\end{aligned}$$

and we see that the choice of V is driven, as in the case of regular histograms, by the quantity

$$(V-1)^{-1}Cm^*+(V-1)\Delta_{m^*}. \tag{29}$$

The same arguments as before show that the optimal choice of V then depends on the ratio m^*/Δ_{m^*} and therefore on s in many situations. This dependence of the optimal value of V with respect to the true density s will actually be confirmed by our simulations below. A density which is difficult to estimate by a histogram with a few bins will lead to a large value of m^* hence a large optimal V while a simple density, for which m^* is rather small, is better estimated by a V-fold with a small V. In the case of a finite set $\mathcal{M}$, which is the practical one, and of equal weights, which is the simplest but suboptimal choice, the optimal V varies like m^*.

3.3 Kernel estimators

We consider here estimation of an unknown density s by a kernel estimator $\widehat{s}_w$ using a non-negative kernel K and a positive bandwidth w which means that

$$\widehat{s}_w(x)=\sum_{i=1}^n K_w(x-X_i)\quad\text{with}\quad K_w(y)=w^{-1}K\left(w^{-1}y\right). \tag{30}$$

Although there are many papers which study the performance of kernel estimators, in particular their risk with respect to $\mathbb{L}_p$-type losses, we were unable to find a result about their non-asymptotic risk with respect to the squared Hellinger loss. This is why we provide one below, the proof of which is deferred to Section 6.

Theorem 2. *Let s be a density on the real line which is supported on an interval of length $2L$ and such that $\sqrt{s}$ has an $\mathbb{L}_2$-modulus of continuity*

$$\omega_2\left(\sqrt{s},\eta\right)=\sup_{|z|\leq\eta}\left\|\sqrt{s}\left(\cdot+z\right)-\sqrt{s}\right\|=\sqrt{2}\sup_{|z|\leq\eta}h\left(s(\cdot+z),s\right). \tag{31}$$

Let ϕ be a nondecreasing and concave function on $[0,+\infty)$ with $\phi(0)=0$ and $\omega_2\left(\sqrt{s},\eta\right)\leq\phi(\eta)$ for $\eta\geq 0$. Assume moreover that the kernel K is bounded with $\int x^2K(x)\,dx<+\infty$ and that it is ultimately monotone around $-\infty$ and $+\infty$. Then the kernel estimator $\widehat{s}_w$ given by (30) satisfies

$$\mathbb{E}_s\left[h^2\left(\widehat{s}_w,s\right)\right]\leq 2\left[\int_{\mathbb{R}}\left(1\vee x^2\right)K(x)\,dx\right]\phi^2(w)+\frac{2L\|K\|_\infty}{nw}+\frac{C(K)}{n}, \tag{32}$$

where the constant C only depends on the kernel K and is equal to 1 when K is unimodal.

If we restrict to densities s with a known compact support, this bound takes the form (28) with the choice $m = w^{-1} + 1$. A "classical" smoothness assumption on $\sqrt{s}$ corresponds to the choice $\phi(\eta) = M\eta^\alpha$, for some exponent $\alpha \in (0, 1]$. In this case the smallest quantity M such that $\omega_2(\sqrt{s}, \eta) \leq \phi(\eta)$ holds true is merely the Besov semi-norm of $\sqrt{s}$ in the Besov space $B^\alpha_{2,\infty}$. In such a case, we see that the optimal value of η is of order $n^{-1/(2\alpha+1)}$, leading to a risk bound of order $n^{-2\alpha/(2\alpha+1)}$. This is completely analogous to what we get for the squared $\mathbb{L}_2$-risk, apart from the fact that for Hellinger we put the smoothness assumption on $\sqrt{s}$ instead of s.

3.4 Handling arbitrary estimators

The previous construction of TVF-estimators is only valid for genuine preliminary density estimators $\widehat{s}_m$, that is such that $\widehat{s}_m(x) \geq 0$ for all $x \in \mathcal{X}$ and $\int \widehat{s}_m(x)\, d\mu(x) = 1$, but this is definitely not the case for all classical estimators. For instance additive estimators given by (12) do not satisfy these requirements when the function $\mathcal{K}$ may take negative values. This actually happens for projection estimators derived from wavelet expansions or kernel estimators based on kernels that take negative values. Not only TVF-estimators cannot be built from preliminary estimators that take negative values but the Hellinger distance cannot be defined for such estimators since it involves the square roots of the densities. There is actually a simple and reasonable solution to this problem which is to transform any function t such that $\int_{t>0} t\, d\mu > 0$ into a probability density $\pi(t)$ with respect to μ using the following operator π:

$$\pi(t) = \frac{t \vee 0}{\int (t(x) \vee 0)\, d\mu(x)}. \tag{33}$$

It is clear that for any probability density s, $|s(x) - (t(x) \vee 0)| \leq |s(x) - t(x)|$ so that $t \vee 0$ is closer from s than t for any reasonable distance, including all $\mathbb{L}_p$-distances. Moreover the following lemma shows that $h(s, \pi(t)) \leq \left\| \sqrt{s} - \sqrt{t \vee 0} \right\|$ which justifies the use of the transformation π when dealing with the Hellinger distance.

Lemma 3. *Let f, g be two nonegative elements of $\mathbb{L}_2(\mu)$ with $\|f\| = 1$ and $\|g\| > 0$. Let $\overline{g} = g/\|g\|$ so that $\|\overline{g}\| = 1$. Then*

$$\|f - \overline{g}\|^2 \leq \frac{4\|f - g\|^2}{4 - \|f - \overline{g}\|^2} \leq 2\|f - g\|^2.$$

If s is a density with respect to μ, g a nonegative element of $\mathbb{L}_2(\mu)$ with positive norm and $u = (g/\|g\|)^2$, then u is also a density with respect to μ and

$$h^2(s, u) \leq 1 - \sqrt{1 - \left(\left\| \sqrt{s} - g \right\|^2 \wedge 1 \right)} \leq \left\| \sqrt{s} - g \right\|^2 \wedge 1.$$

If, in particular, t is an arbitrary element of $\mathbb{L}_1(\mu)$ such that $\int (t \vee 0)\, d\mu > 0$, then

$$h^2(s, \pi(t)) \leq 1 - \sqrt{1 - \left(\left\| \sqrt{s} - \sqrt{t \vee 0} \right\|^2 \wedge 1 \right)} \leq \left\| \sqrt{s} - \sqrt{t \vee 0} \right\|^2 \wedge 1.$$

Proof. Let $\|g\| = \lambda$ so that $g = \lambda \overline{g}$ and let $\rho = \langle f, \overline{g} \rangle \in [0, 1]$. Then

$$\|f - g\|^2 = 1 + \lambda^2 - 2\lambda\rho \qquad \text{and} \qquad \|f - \overline{g}\|^2 = 2(1 - \rho) \leq 2.$$

It follows that, for a given value of ρ, the minimum value of $\|f-g\|^2$ is obtained for $\lambda=\rho$ and equal to $1-\rho^2$ which implies that

$$\left(\frac{\|f-\overline{g}\|}{\|f-g\|}\right)^2 \leq \frac{2}{1+\rho} = \frac{4}{4-\|f-\overline{g}\|^2} \leq 2. \tag{34}$$

If $f=\sqrt{s}$, then $\rho=\rho(s,u)=1-h^2(s,u)$ and $\|f-\overline{g}\|^2=2h^2(s,u)$, so that (34) becomes

$$h^2(s,u) \leq \frac{\|f-g\|^2}{1+\rho(s,u)} = \frac{\|f-g\|^2}{2-h^2(s,u)}$$

and, since $h(s,u)\leq 1$, it also follows from elementary calculus that

$$h^2(s,u) \leq 1-\sqrt{1-(\|f-g\|^2\wedge 1)} \leq \|f-g\|^2\wedge 1.$$

The last inequality is just the case of $g=\sqrt{t\vee 0}$. □

Using the transformation π amounts to replace the initial family $\{\mathcal{A}_m,\, m\in\mathcal{M}\}$ by a new one $\{\mathcal{A}'_m,\, m\in\mathcal{M}\}$ via the tranformation $\mathcal{A}'_m(\mathbf{Y}_k)=\pi\left(\mathcal{A}_m(\mathbf{Y}_k)\right)$ which results in procedures that now make sense for the Hellinger loss. Unfortunately, this transformation does not preserve the linearity so that if we apply this recipe to projection or kernel estimators, we cannot know whether the transformed estimators satisfy Assumption (LOSS). Nevertheless, as we have seen in Section 2.5, we may change the definition of TVF-estimators to (17) in order to solve this problem.

Starting from a family of estimators that are not probability densities, we merely begin with a preliminary application of the transformation π, as given by (33), and then define our modified TVF-estimator via (17) so that Theorem 1 applies to the family of procedures $\{\mathcal{A}'_m,\, m\in\mathcal{M}\}$.

4 Empirical study

The theoretical bounds that we have derived, for instance (20), are quite pessimistic because of the large constants that are present in our risk bounds. It is therefore crucial to know whether such large values are only artifacts or really enter the risk. In order to check the real quality of our selection procedure and evaluate the influence of the various parameters involved in it, we performed an extensive set of simulations the results of which are summarized below.

4.1 Simulation protocol

We studied the performances of the TVF procedure on 18 out of the 28 densities described in the *benchden* [2] R-package [17] which provides a full implementation of the distributions introduced in [4] as benchmarks for nonparametric density estimation. We only show our simulations for the eleven densities in the subset $\mathcal{L}=\{s_i,\ i=1,2,3,4,5,7,12,13,22,23,24\}$ (where the indices refer to the list of benchden) the graphs of which are shown in Figure 1, except for the uniform density s_1 on $[0,1]$. For a given loss $\ell=h^2,d_1$ or d_2^2 (respectively the squared Hellinger, $\mathbb{L}_1$- and squared $\mathbb{L}_2$-losses), we decided to evaluate the accuracy of

[2] Available on the CRAN http://cran.r-project.org.

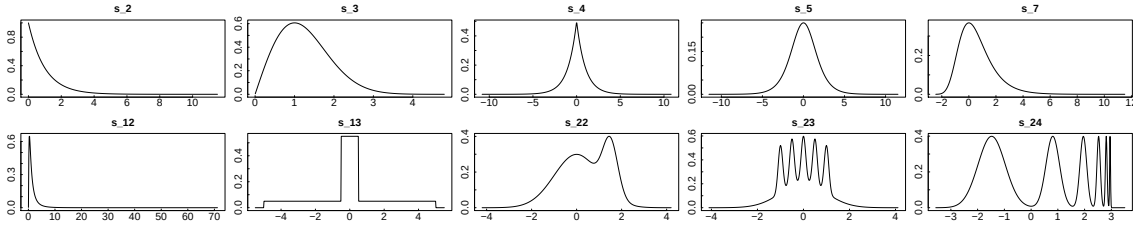

Figure 1: Graphs of all densities mentioned in the paper apart from the uniform.

some estimator $\widetilde{s} = \widehat{s}_{\widehat{m}}$ by empirically estimating its risk $R(\widetilde{s}, s, \ell) = \mathbb{E}_s\left[\ell(s, \widetilde{s})\right]$. To do so, we generated 1000 pseudo-random samples $\mathbf{X}^i = \{X_1^i, \ldots, X_n^i\}$, $1 \leq i \leq 1000$, of size n and density s and approximated $R(\widetilde{s}, s, \ell)$ by its empirical version:

$$\overline{R}_n\left(\widetilde{s}, s, \ell\right) = \frac{1}{1000} \sum_{i=1}^{1000} \ell\left(s, \widetilde{s}(\mathbf{X}^i)\right).$$

As in [14], we considered several families of estimators. We present here our simulations for the well-known problems of bandwidth selection for kernel estimators with a Gaussian kernel and the choice of the bin number for regular histograms. We therefore introduce the following families of estimators.

- $\mathcal{F}_{\mathrm{R}}$ is the set of regular histograms with bin number varying from 1 to $\lceil n/\log n \rceil$ as described in [10],
- $\mathcal{F}_{\mathrm{K}}$ is the set of Gaussian kernel estimators with bandwidths w_m of the form

$$w_m = \frac{1}{n \log n}\left(1 + \frac{1.5}{\log n}\right)^m, \qquad \text{for } m = 1, \ldots, (\log n)^2,$$

- $\mathcal{F}_{\mathrm{KR}} = \mathcal{F}_{\mathrm{K}} \cup \mathcal{F}_{\mathrm{R}}$.

Besides the classical VF methods, we considered two alternative procedures that are known to perform well in practice in order to have an idea of the performance of the T-V-fold as compared to some especially calibrated methods. When studying the problem of bandwidth selection, we compared the TVF with the unbiased cross-validation selector, implemented in the *density* generic function available in R, which provides an estimator which does not belong to the set $\{\widehat{s}_m,\, m \in \mathcal{M}\}$. When dealing with the bin number choice we implemented the penalization procedure of Birgé and Rozenholc (described in [10]) which selects a regular histogram in $\mathcal{F}_{\mathrm{R}}$. These two competitors will be denoted "UCV" and "BR" respectively in our study. To implement the TVF and process our simulations we used an algorithm which is described in Section 5 with the tests defined in (9) and constant weights $\Delta_m = \Delta = 0$ for all $m \in \mathcal{M}$.

We made thousands of simulations (varying the sample size n, the density, the family of estimators, the number V of splits in our V-fold procedures, etc.) but since the results we found were very similar in all situations, we only show the conclusion for $n = 500$ and $V = 2, 5, 10$ and 20.

4.2 The influence of the parameter θ

As in [14, Section 5.1], we have studied the influence on the performance of the TVF procedure of the parameter θ that is used in the definition of the test statistic (9). The parameter influences the performance of the tests $\psi_{t,u}$ as shown by (7) and (8) and therefore the whole procedure. Since on the one hand $\theta = 0$ corresponds to the KLVF and on the other hand θ must be less than $1/2$, we made comparisons between the versions of $\widetilde{s}_{\mathrm{TVF}}$ deduced from the tests with $\theta \in \Theta = \{1/16, 1/8, 1/4, 3/8, 7/16\}$. For the sake of clarity and to emphasize the stability of the behavior of the procedure in terms of risk, we present for each V the ratio

$$\inf_{s\in\mathcal{L}} \left\{ \inf_{\theta\in\Theta} \overline{R}_n \left(\widehat{s}_{\widehat{m}(\theta)}, s, h^2 \right) \Big/ \sup_{\theta\in\Theta} \overline{R}_n \left(\widehat{s}_{\widehat{m}(\theta)}, s, h^2 \right) \right\}, \tag{35}$$

which gives the largest difference in terms of risk among the densities in $\mathcal{L}$. The closer the ratio to 1, the more stable the procedure with respect to the variations of θ. We may conclude

family	$V = 2$	$V = 5$	$V = 10$	$V = 20$
$\mathcal{F}_{\mathrm{R}}$	92,95	94,87	96,39	96,96
$\mathcal{F}_{\mathrm{K}}$	91,31	92,94	94,79	96,44
$\mathcal{F}_{\mathrm{KR}}$	87,81	94,36	97,48	95,15

Table 1: 100 times the ratio (35) for $n = 500$ and families $\mathcal{F}_{\mathrm{R}}$, $\mathcal{F}_{\mathrm{K}}$ and $\mathcal{F}_{\mathrm{KR}}$.

from this picture that θ has little influence on the quality of the resulting estimator for families $\mathcal{F}_{\mathrm{K}}$ and $\mathcal{F}_{\mathrm{R}}$, even if we did observe that $\theta = 1/16$ is in general slightly worse than the other values (in particular for the family $\mathcal{F}_{\mathrm{R}}$). Considering family $\mathcal{F}_{\mathrm{KR}}$, we have observed that there might be some noticeable difference for $V = 2$ for one specific density. Nevertheless no clear conclusion can be derived from our simulations as the best value of θ varies with the setting. Finally, it appears that the choice $\theta = 1/4$ is always satisfactory and should be recommended.

4.3 About the choice of V

The main question when considering VF type procedures is maybe "which V is optimal?" or, more generally, "what is the influence of V on the quality of the VF procedure?". According to our theoretical study in Section 3 the optimal value of V depends on the optimal value m^* of m. In the case of equal weights the best V appears to be an increasing function of m^*. In the case of histograms, if the best one has many bins, one should take a large value of V and the same would hold for a kernel estimator with a small bandwidth. To understand what actually happens in practice, we study here how the risk of the chosen estimator behaves when V varies.

Since θ has little influence, we made all the simulations with $\theta = 1/4$. We also implemented the calibrated procedures BR and UCV described in Section 4.1 in order to have a benchmark for the risk for the families $\mathcal{F}_{\mathrm{R}}$ and $\mathcal{F}_{\mathrm{K}}$ respectively.

The empirical results summarized in Table 2 actually confirm what we derived from (29). The quality of the estimation increases with V when the true density is difficult to estimate which corresponds to an optimal estimator $\widehat{s}_{m^*}$ with a large value of m^* in (29). For a simple density like the uniform s_1 which is better estimated by an histogram with few bins, the best choice of V is 2 for the families $\mathcal{F}_{\mathrm{R}}$ and $\mathcal{F}_{\mathrm{KR}}$ which include histograms. On the contrary, when dealing with the family $\mathcal{F}_{\mathrm{K}}$ for which s_1 is not easy to estimate, we need to use a larger

family	V	s_1	s_2	s_3	s_4	s_5	s_7	s_{12}	s_{13}	s_{22}	s_{23}	s_{24}
$\mathcal{F}_{\mathrm{R}}$	2	**2,9**	10,4	9,29	13,8	10,9	11,4	17,9	14,5	10,5	20,8	27,5
	5	4,31	9,9	8,75	12,7	10	10,6	17,3	**13,5**	9,56	18,4	25,2
	10	6,18	9,81	8,64	12,3	9,77	10,6	**17,2**	13,7	9,51	**17,8**	**24,8**
	20	9,39	**9,65**	**8,54**	**12,2**	**9,59**	**10,4**	17,3	14,1	**9,28**	17,9	**24,8**
	BR	2,20	9,94	9,27	12,98	10,53	11,14	17,85	14,63	10,37	17,98	25,15
$\mathcal{F}_{\mathrm{K}}$	2	15,4	29,9	5,67	5,1	**3,56**	4,26	28,5	20	3,96	10,6	18,1
	5	12,7	25,5	5,06	**4,95**	3,61	**3,98**	23,4	18,1	**3,86**	9,28	16,2
	10	12,4	24,3	**4,94**	5,01	3,96	4,04	21,8	17,7	3,91	9,08	15,8
	20	**12,2**	**23,5**	4,97	5,41	4,9	4,27	**20,9**	**17,6**	4,11	**9,05**	**15,7**
	UCV	15,86	22,20	5,57	6,16	3,74	4,10	18,80	17,16	3,88	9,52	15,91
$\mathcal{F}_{\mathrm{KR}}$	2	**2,88**	10,4	8,32	6,35	5,81	6,57	18,5	14,4	7,3	12,8	20
	5	4	9,91	7,86	**5,64**	**5,11**	**6,06**	17,7	**13,2**	**5,76**	9,66	16,7
	10	4,34	9,95	7,66	**5,64**	5,4	6,18	17,6	13,7	5,82	9,12	16
	20	4,34	**9,86**	**7,49**	5,91	5,81	6,5	**17,5**	14,5	5,88	**9,08**	**15,7**

Table 2: 10^3 times the Hellinger risks of the TVF procedure.

value of V. A similar situation occurs with densities s_4, s_5, s_7 and s_{22} which appears to be easily estimated by a kernel estimator with a large bandwidth but poorly by histograms. It seems that, apart from the exceptional situation of s_1, the best value of V is not 2 and the most significant gain appears between $V = 2$ and $V = 5$, then the quality sometimes keeps improving from $V = 5$ to $V = 20$, but with very little difference between $V = 10$ and $V = 20$.

Interestingly, we also observe that when using the mixed collection $\mathcal{F}_{\mathrm{KR}}$ the TVF procedure shows a good adaptation behaviour since it selects the best family in all settings. For instance for s_5 it chooses a kernel estimator since these are better than histograms for estimating it, whereas it selects an histogram for s_2 for the opposite reason.

The numerical complexity of the TVF procedure is quite important in practice and increases with V so that large values of V should be avoided because they lead to a much larger computation time. In particular the Leave-one-out ($V = n$) should be excluded since it is typically impossible to compute it in a reasonable amount of time. Of course, since the optimal value of V, as we have seen, depends on unknown properties of the procedures with respect to the true density it is not possible to practically define an optimal choice of V. Nevertheless our empirical study suggests that a good compromise, which leads to both a reasonable computation time and a good performance (apart from some exceptional situations like the estimation of the uniform by histograms), is $V = 5$. We would therefore recommend the user to process the TVF procedure with this value.

4.4 Comparison with others VF procedures

The goal of this section is to compare our TVF procedure with other general VF procedures namely LSVF and KLVF, which do not depend on the family of estimators from which we estimate s. In order to compare two VF procedures $\tilde{t}_1$ and $\tilde{t}_2$, we consider the $\log_2$-ratio of their empirical risk,

$$\overline{W}_s\left(\tilde{t}_1, \tilde{t}_2, \ell\right) = \log_2 \frac{\overline{R}_n\left(\tilde{t}_1, s, \ell\right)}{\overline{R}_n\left(\tilde{t}_2, s, \ell\right)}.$$

Thus $\overline{W}_s\left(\tilde{t}_1, \tilde{t}_2, \ell\right) = c$ means that $\overline{R}_n\left(\tilde{t}_1, s, \ell\right) = 2^c \times \overline{R}_n\left(\tilde{t}_2, s, \ell\right)$. Hence, for a given density s, $\tilde{t}_2$ is a better estimator than $\tilde{t}_1$ if $c > 0$. In our empirical study, a selection procedure $\tilde{t}_2$

is thus considered better than $\tilde{t}_1$ in terms of risk for a given loss function ℓ if the values of $\overline{W}_s(\tilde{t}_1, \tilde{t}_2, \ell)$ are positive when the density s varies in $\mathcal{L}$.

Rather than presenting exhaustive results, that is the evaluation of $\overline{W}_s$ for all densities s in $\mathcal{L}$, different loss functions, various observations numbers n and different choices of V, we shall illustrate the results of our simulations by showing boxplots of $\{\overline{W}_s(\tilde{t}_1, \tilde{t}_2, \ell),\ s \in \mathcal{L}\}$ with the discriminating value zero emphasized in red. We actually observed similar results and behaviours for all losses and all sample sizes and therefore present here only the results for $\ell = h^2$ and $n = 500$ for the sake of simplicity. Figure 2 is built with $\tilde{t}_1 = \widehat{s}_{\widehat{m}_{\mathrm{LSVF}}}$ (upper line) or $\widehat{s}_{\widehat{m}_{\mathrm{KLVF}}}$ (bottom line) and $\tilde{t}_2 = \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}$ with $\theta = 1/4$.

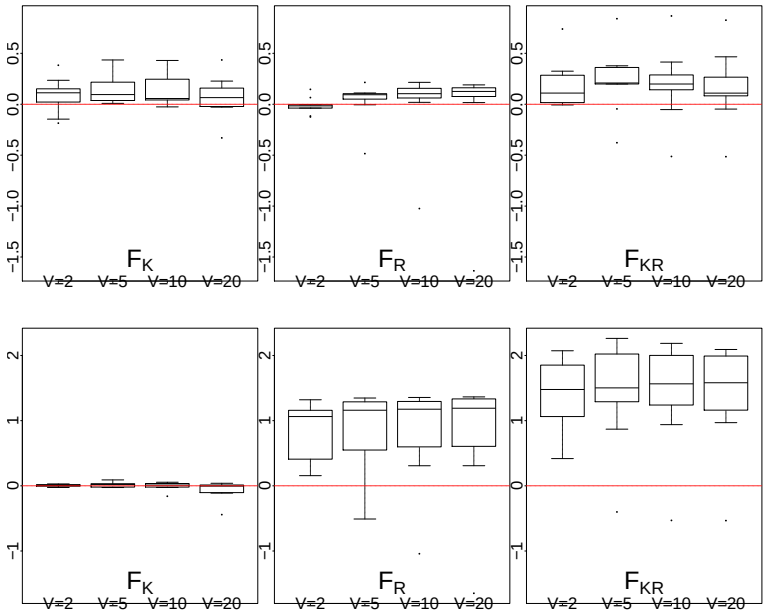

Figure 2: From left to right, the boxplots $\overline{W}_s(\tilde{s}, \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}, h^2)$ using families $\mathcal{F}_{\mathrm{K}}, \mathcal{F}_{\mathrm{R}}, \mathcal{F}_{\mathrm{KR}}$ (up for $\tilde{s} = \widehat{s}_{\widehat{m}_{\mathrm{LSVF}}}$, down for $\tilde{s} = \widehat{s}_{\widehat{m}_{\mathrm{KLVF}}}$). Each subfigure shows the boxplots for $V = 2, 5, 10$ and 20. The horizontal red dotted line indicates the reference value 0.

In nearly all cases, the median and most of the distribution are positive, which means that the TVF outperforms LSVF (with an average gain of about 20% for the three families of estimators $\mathcal{F}_{\mathrm{K}}$, $\mathcal{F}_{\mathrm{R}}$ and $\mathcal{F}_{\mathrm{KR}}$) and KLVF as well. For the collection $\mathcal{F}_{\mathrm{K}}$ we observe that the empirical risks of TVF and KLVF are similar with boxplots of $\overline{W}_s(\widehat{s}_{\widehat{m}_{\mathrm{KLVF}}}, \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}, h^2)$ highly concentrated around zero. But there is a huge difference between TVF and KLVF procedures for families $\mathcal{F}_{\mathrm{R}}$ and $\mathcal{F}_{\mathrm{KR}}$ (average gain of about 100% and 180% respectively). For the uniform density estimated with regular histograms, the estimator derived from our procedure is worse since we found, for both classical VF, $\overline{W}_{s_1}(\tilde{s}, \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}, h^2) < 0$ (with an increasing difference with V for $\mathcal{F}_{\mathrm{R}}$). Finally, let us notice that the difference between TVF and classical VF does not change much with V.

5 Our computational algorithm

For the practical computation of the TVF as well as any other VF procedure, we assume that $\mathcal{M}$ is finite with cardinality M.

Let us compare the complexity of a classical V-fold method with ours. Since for every VF method the construction of all partial estimators $(\widehat{s}_{m,j})_{1\leq j\leq V, 1\leq m\leq M}$ is required, we only have to focus on the "validation part" which requires to compute all quantities $\mathcal{D}_j^2(m)$ for $1 \leq j \leq V$ and $m \in \mathcal{M}$ and therefore to perform all tests $\psi_{l,m}(\mathbf{X}_j)$ for $1 \leq j \leq V$ and $l, m \in \mathcal{M}$ with $l \neq m$. This means performing $V \times M \times (M-1)/2$ tests leading to a computational cost of order $O(V \times M^2)$ that can be prohibitive as compared to the one of either LSVF or KLVF which have a maximum complexity of order $O(V \times M)$ (since in this case no more than M calculations are needed for each split). For instance, a 10-fold with 100 different procedures would require at most 1000 evaluations for a classical VF whereas we would need the computation of 49500 tests for the TVF. It is already huge and does not even take into account the computation of the distances $h^2(\widehat{s}_{l,j}, \widehat{s}_{m,j})$, each one requiring the evaluation of an integral. Therefore a "naive" algorithm based on the computation of all the $V \times M$ values $\mathcal{D}_j^2(m)$ would be very slow.

Fortunately, there is a smarter way to determine which $\widehat{m}$ minimizes $\overline{\mathcal{D}}(\cdot)$ over $\mathcal{M}$. Our algorithm is inspired in some way by the one described in [14, Section 3]. In order to explain how this "fast" algorithm works, it will be convenient to single an element of $\mathcal{M}$, that we shall denote by "m_s", to serve as a starting point for our algorithm which begins with the computation of $\overline{\mathcal{D}}(m_s)$. We store in R the minimal value of those $\overline{\mathcal{D}}^2(m)$ that have already been computed and in *opt* the corresponding optimal value of m with initial values $opt = m_s$ and $R = \overline{\mathcal{D}}^2(m_s)$. We update them after each computation of a new $\overline{\mathcal{D}}^2(m)$ such that $\overline{\mathcal{D}}^2(m) < R$, then setting $opt := m$ and $R := \overline{\mathcal{D}}^2(opt)$ so that R can only decrease during the computational procedure.

By (11), minimizing $\overline{\mathcal{D}}^2(m)$ is equivalent to minimizing $\sum_{j=1}^V \mathcal{D}_j^2(m)$. Since

$$\mathcal{D}_j^2(m) = \sup_{l\in\mathcal{M}_m} h^2\left(\widehat{s}_{l,j}, \widehat{s}_{m,j}\right) \mathbb{1}_{\left\{\psi_{l,m}(\mathbf{X})=l\right\}} \quad \text{with} \quad \mathcal{M}_m = \mathcal{M} \setminus \{m\},$$

one can compute it iteratively, starting with $\mathcal{L}_j(m) = 0$ and setting

$$\mathcal{L}_j(m) := \max\left(\mathcal{L}_j(m), h^2\left(\widehat{s}_{l,j}, \widehat{s}_{m,j}\right)\right) \quad \text{when} \quad \psi_{l,m}(\mathbf{X}_j) = l \quad \text{for} \quad l \in \mathcal{M}_m.$$

If $\psi_{l,m}(\mathbf{X}_j) = m$ we can instead update $\mathcal{L}_j(l)$ by $\mathcal{L}_j(l) := \max(\mathcal{L}_j(l), h^2(\widehat{s}_{l,j}, \widehat{s}_{m,j}))$ using the result of the test $\psi_{l,m}(\mathbf{X}_j)$ for the calculation of both $\mathcal{D}_j^2(m)$ and $\mathcal{D}_j^2(l)$. Our algorithm proceeds in this way, with a set of M V-dimensional vectors $\mathcal{L}_\cdot(m)$, $m \in \mathcal{M}$, initially set to zero. The updating procedure of $\mathcal{L}_j(m)$ stops when all updates, with $l \in \mathcal{M}_m$, have been done (which means that the present value of $\mathcal{L}_j(m)$ is $\mathcal{D}_j^2(m)$) and we finally set $\overline{\mathcal{D}}^2(m) = \sum_{j=1}^V \mathcal{L}_j(m)$.

We also use another trick in order to shorten our computations. Since $\mathcal{L}_j(m)$ can only increase during the updating procedure, $\sum_{j=1}^V \mathcal{L}_j(m)$ is, at any time, a lower bound for $\overline{\mathcal{D}}^2(m)$, whatever $m \in \mathcal{M}$. Therefore it is useless to go on with the computation of the vector $\mathcal{L}_\cdot(m)$ if $\sum_{j=1}^V \mathcal{L}_j(m) > R$ since then $\overline{\mathcal{D}}^2(m) \geq \sum_{j=1}^V \mathcal{L}_j(m)$ cannot minimize the function $\overline{\mathcal{D}}(\cdot)$ over $\mathcal{M}$. Taking this fact into account, we denote by $\mathcal{G} \subset \mathcal{M}$ the set of all procedures which are potentially "better" than the current optimal one stored in *opt*. This means that we store in

$\mathcal{G}$ all $m \in \mathcal{M}$ for which we do not yet know whether $\overline{\mathcal{D}}^2(m) < R$ or not and each time we find m such that $\sum_{j=1}^{V} \mathcal{L}_j(m) > R$, we remove it from $\mathcal{G}$. We also remove m from $\mathcal{G}$ once we have computed $\overline{\mathcal{D}}^2(m)$ with $m \in \mathcal{G}$ and then proceed with the computation of some new vector $\mathcal{L}.(l)$ for $l \in \mathcal{G}$ until $\mathcal{G}$ is empty and the algorithm stops with the final value $\widehat{m} = opt$.

Algorithm 1: Selection of the TVF estimator

Initialization:
Set $\mathcal{G} = \mathcal{M}_{m_s}$ and $opt = m_s$
for $(l \in \mathcal{M})$ **do**
  for $(j = 1, \ldots, V)$ **do**
    $\mathcal{L}_j(l) = 0$
  end
end

1**st step**:
for $(l \in \mathcal{G})$ **do**
  Compute $\psi_{m_s,l}(\mathbf{X}_j)$
  if $(\psi_{m_s,l}(\mathbf{X}_j) = m_s)$ **then**
    $\mathcal{L}_j(l) = h^2(\widehat{s}_{l,j}, \widehat{s}_{m_s,j})$
  else
    $\mathcal{L}_j(m_s) = \max(\mathcal{L}_j(m_s), h^2(\widehat{s}_{l,j}, \widehat{s}_{m_s,j}))$
  end
end
Set $R = \sum_{j=1}^{V} \mathcal{L}_j(m_s)$ and $\mathcal{G} = \mathcal{G} \setminus \{l \in \mathcal{G} : \sum_{j=1}^{V} \mathcal{L}_j(l) > R\}$

Next steps:
while $(|\mathcal{G}| > 0)$ **do**
  Choose $m \in \mathcal{G}$ and set $\mathcal{G} = \mathcal{G} \setminus \{m\}$
  for $(j = 1, \ldots, V)$ **do**
    for $(l \in \mathcal{M}_m)$ **do**
      Compute $\psi_{m,l}(\mathbf{X}_j)$ `// if it has not been done yet`
      if $(\psi_{m,l}(\mathbf{X}_j) = m$ *and* $l \in \mathcal{G})$ **then**
        $\mathcal{L}_j(l) = \max(\mathcal{L}_j(l), h^2(\widehat{s}_{l,j}, \widehat{s}_{m,j}))$
        if $(\sum_{i=1}^{V} \mathcal{L}_i(l) > R)$ **then**
          $\mathcal{G} = \mathcal{G} \setminus \{l\}$
        end
      end
      if $(\psi_{m,l}(\mathbf{X}_j) = l)$ **then**
        $\mathcal{L}_j(m) = \max(\mathcal{L}_j(m), h^2(\widehat{s}_{l,j}, \widehat{s}_{m,j}))$
        if $(\sum_{i=1}^{V} \mathcal{L}_i(m) > R)$ **then**
          break `// quit the two "for" loops`
        end
      end
    end
  end
  if $(\sum_{j=1}^{V} \mathcal{L}_j(m) < R)$ **then**
    Set $opt = m$, $R = \sum_{j=1}^{V} \mathcal{L}_j(m)$ and $\mathcal{G} = \mathcal{G} \setminus \{l \in \mathcal{G} : \sum_{j=1}^{V} \mathcal{L}_j(l) > R\}$
  end
end
Return *opt*

Some important remarks

- The algorithm is designed to work with any test procedure ψ which satisfies Assumption

(TEST) or, more generally, Assumption (TEST'), like the procedures based on the statistics (9) or (18).

- It is important to notice that, at any step, we cannot "delete" once and for all the procedures which do not belong to the set $\mathcal{G}$. Even if we do not compute the value of $\overline{\mathcal{D}}$ for these procedures, we still need to test them against the remaining procedures in $\mathcal{G}$.
- We hoped that by starting from a good initial estimator, only a few procedures would be in the first set $\mathcal{G}$, resulting in just a few tests. In the simulations we always started from $m_s = \widehat{m}_{\mathrm{LSVF}}$ since the computation of $\widehat{m}_{\mathrm{LSVF}}$ is less costly than the one of $\widehat{m}_{\mathrm{TVF}}$ and provides a good starting point. If $\overline{\mathcal{D}}(\widehat{m}_{\mathrm{LSVF}}) = 0$ at the first step the algorithm stops immediately and the chosen procedure is $\widehat{m} = \widehat{m}_{\mathrm{LSVF}}$. In this special case, the complexity of our algorithm is the same as the one of the classical approach.
- Clearly, the choice of m at line 17 of the algorithm, as well as the choice of the starting procedure, have no influence on the final estimator, only on the computational time. To avoid a quadratic complexity, we need to ensure that we don't "jump" to the worst procedure inside the set $\mathcal{G}$ at each iteration. In our simulations, we chose to jump to the statistical method $k \in \mathcal{G}$ with the lowest temporary criterion among the procedures in $\mathcal{G}$, that is $k = \mathrm{argmin}_{l\in\mathcal{G}} \sum_{j=1}^{V} \mathcal{L}_j(l)$. We also tried two alternative options: jumping to $k = \mathrm{argmax}_{l\in\mathcal{G}} \sum_{j=1}^{V} \mathcal{L}_j(l)$ and to the most chosen statistical method k in $\mathcal{G}$ against m. Both options lead of course to the same final estimator but were definitely slower.

6 Proof of Theorem 2

First note that the kernel estimator $\widehat{s}_w$ can be written, according to (30), $K_w * P_n$ where $P_n = n^{-1}\sum_{i=1}^{n} \delta_{X_i}$ denotes the empirical measure based on the i.i.d. sample $X_1, ..., X_n$. It then follows from the triangle inequality that

$$\begin{aligned} \mathbb{E}_s\left[h^2\left(\widehat{s}_w, s\right)\right] &= \frac{1}{2}\mathbb{E}_s\left\|\sqrt{K_w * P_n} - \sqrt{s}\right\|^2 \\ &\leq \left\|\sqrt{s} - \sqrt{K_w * s}\right\|^2 + \mathbb{E}_s\left\|\sqrt{K_w * P_n} - \sqrt{K_w * s}\right\|^2, \end{aligned} \tag{36}$$

which is the usual bound of the risk as squared bias plus variance, and we shall bound both terms successively.

6.1 Bounding the bias

It is well known that whenever the function s belongs to $\mathbb{L}_2$, the quality of approximation of s by the convolution $K_w * s$ depends on the modulus of continuity $\omega_2(s, \cdot)$ of s in $\mathbb{L}_2$ as given by (31). If we consider the Hellinger distance instead of the $\mathbb{L}_2$-distance it is expected that the quality of approximation should rather depend on the the modulus of continuity of $\sqrt{s}$ instead. The control of the bias term is provided by the following lemma:

Lemma 4. *Let s and K be some density functions with respect to Lebesgue measure on the real line. Let $g = \sqrt{s}$ and assume that $\omega_2(g, \eta) \leq \phi(\eta)$ for every nonnegative η and some nondecreasing and concave function ϕ on $[0, \infty)$ with $\phi(0) = 0$. Then for every positive real number w*

$$\left\|\sqrt{s} - \sqrt{K_w * s}\right\|^2 \leq 2\left[\int_{\mathbb{R}} \left(1 \vee x^2\right) K(x)\, dx\right] \phi^2(w). \tag{37}$$

Proof. The key of the proof is to compare $D^2 = \left\| g - \sqrt{K_w * g^2} \right\|^2$ with $\Delta^2 = \|g - K_w * g\|^2$. Our arguments are easier to explain within a probabilistic framework. Let ξ be some random variable with density K with respect to the Lebesgue measure. Then the convolution operator can be written as

$$(K_w * g)(x) = \mathbb{E}[g(x - w\xi)] \text{ for all } x \in \mathbb{R}.$$

From Jensen's inequality we know that

$$\mathbb{E}[g(x - w\xi)] \leq \sqrt{\mathbb{E}\left[g^2(x - w\xi)\right]},$$

or equivalently $\sqrt{K_w * g^2} \geq K_w * g$, and a fortiori,

$$\int g(x)\sqrt{(K_w * g^2)(x)}\, dx \geq \int g(x)(K_w * g)(x)\, dx. \tag{38}$$

Expanding the square norms, we derive from (38) that

$$D^2 - \Delta^2 \leq \left\| \sqrt{K_w * g^2} \right\|^2 - \|K_w * g\|^2 .$$

The trick is to notice that

$$\left\| \sqrt{K_w * g^2} \right\|^2 - \|K_w * g\|^2 = \int_{\mathbb{R}} \mathrm{Var}\left(g(x - w\xi)\right) dx.$$

Now since the computation of the variance is not sensitive to the substraction of a constant

$$\mathrm{Var}\left(g(x - w\xi)\right) = \mathrm{Var}\left(g(x - w\xi) - g(x)\right) \leq \mathbb{E}\left[\left(g(x - w\xi) - g(x)\right)^2\right]$$

and Fubini's Theorem implies that

$$\left\| \sqrt{K_w * g^2} \right\|^2 - \|K_w * g\|^2 \leq \mathbb{E}\left[\omega_2^2(g, w|\xi|)\right].$$

This achieves the first step of the proof. The second step is straightforward. We just have to bound Δ^2 which is an easy task since

$$\Delta^2 = \int_{\mathbb{R}} \left(\mathbb{E}[g(x - w\xi) - g(x)]\right)^2 dx$$

implies by Jensen's inequality and Fubini's Theorem that

$$\Delta^2 \leq \mathbb{E}\left[\int_{\mathbb{R}} \left(g(x - w\xi) - g(x)\right)^2 dx\right] \leq \mathbb{E}\left[\omega_2^2(g, w|\xi|)\right].$$

Collecting these bounds, we derive that

$$D^2 \leq 2\mathbb{E}\left[\omega_2^2(g, w|\xi|)\right] \leq 2\mathbb{E}\left[\phi^2(w|\xi|)\right].$$

It remains to decouple w and ξ in the last expression above. This can be done by noticing that the monotonicity and concavity properties of ϕ imply that $\phi(w|\xi|) \leq (1 \vee |\xi|)\phi(w)$ and the result follows. □

6.2 The variance term

We now turn to the analysis of the variance term of the Hellinger risk of a kernel estimator.

Lemma 5. *Let us denote by A_w the Borel set $\{x \in \mathbb{R} \,|\, (K_w * s)(x) > 0\}$, then*

$$\mathbb{E}_s\left[\left\|\sqrt{K_w * P_n} - \sqrt{K_w * s}\right\|^2\right] \leq \frac{1}{nw}\int_{A_w} \frac{((K^2)_w * s)(x)}{(K_w * s)(x)}\,dx. \tag{39}$$

In particular if s is supported on an interval of finite length $2L$, then

$$\mathbb{E}_s\left[\left\|\sqrt{K_w * P_n} - \sqrt{K_w * s}\right\|^2\right] \leq \frac{1}{nw}\int \sup_{-L\leq z\leq L} K\left(\frac{x-z}{w}\right)dx. \tag{40}$$

If the kernel K is bounded, non-decreasing on $(-\infty, M_1)$ and non-increasing on $(M_2, +\infty)$ with $M_1 \leq M_2$, then

$$\begin{aligned}\int \sup_{-L\leq z\leq L} K\left(\frac{x-z}{w}\right)dx & \\ \leq\ & 2L\|K\|_\infty + w\left[(M_2 - M_1)\|K\|_\infty + \int_{-\infty}^{M_1} K(x)\,dx + \int_{M_2}^{\infty} K(x)\,dx\right].\end{aligned} \tag{41}$$

If, in particular, K is unimodal, then

$$\mathbb{E}_s\left[\left\|\sqrt{K_w * P_n} - \sqrt{K_w * s}\right\|^2\right] \leq \frac{2L\|K\|_\infty}{nw} + \frac{1}{n}.$$

Proof. Since, for $u, v \geq 0$,

$$(u-v)^2 = \left(\sqrt{u} - \sqrt{v}\right)^2\left(\sqrt{u} + \sqrt{v}\right)^2 \geq v\left(\sqrt{u} - \sqrt{v}\right)^2,$$

it follows that

$$\int\left(\sqrt{u(x)} - \sqrt{v(x)}\right)^2 dx \leq \int_{v(x)>0} \frac{[u(x)-v(x)]^2}{v(x)}\,dx + \int_{v(x)=0} u(x)\,dx,$$

hence

$$\left\|\sqrt{K_w * P_n} - \sqrt{K_w * s}\right\|^2 \leq \int_{A_w} \frac{\left[(K_w * P_n)(x) - (K_w * s)(x)\right]^2}{(K_w * s)(x)}\,dx + \int_{A_w^c} (K_w * P_n)(x)\,dx.$$

By Fubini and the definition of A_w,

$$\mathbb{E}_s\left[\int_{A_w^c} (K_w * P_n)(x)\,dx\right] = \int_{A_w^c} \mathbb{E}_s\left[(K_w * P_n)(x)\right]dx = \int_{A_w^c} (K_w * s)(x)\,dx = 0.$$

Taking expectations and using Fubini again, we therefore get

$$\begin{aligned}\mathbb{E}_s\left[\left\|\sqrt{K_w * P_n} - \sqrt{K_w * s}\right\|^2\right] &\leq \int_{A_w} \frac{\mathbb{E}_s\left[\left[(K_w * P_n)(x) - K_w * s(x)\right]^2\right]}{(K_w * s(x)}\,dx \\ &= \int_{A_w} \frac{\mathrm{Var}\left((K_w * P_n)(x)\right)}{K_w * s(x)}\,dx\end{aligned}$$

and (39) follows since

$$\mathrm{Var}\Big((K_w * P_n)(x)\Big) = \frac{\mathrm{Var}\Big(K_w(x-X)\Big)}{n} \leq \frac{\mathbb{E}_s\left[\Big(K_w(x-X)\Big)^2\right]}{n} = \frac{((K^2)_w * s)(x)}{nw}. \quad (42)$$

Now observe that if f is supported on $[a, a+2L]$,

$$\Big((K^2)_w * s\Big)(x) = \int_a^{a+2L} \frac{1}{w} K^2\left(\frac{x-z}{w}\right) s(z)\, dz \leq \sup_{a\leq z\leq a+2L} K\left(\frac{x-z}{w}\right) (K_w * s)(x),$$

so that by (39)

$$\mathbb{E}_s\left[\left\|\sqrt{K_w * P_n} - \sqrt{K_w * s}\right\|^2\right] \leq \frac{1}{nw} \int \sup_{a\leq z\leq a+2L} K\left(\frac{x-z}{w}\right) dx.$$

Since

$$\int \sup_{a\leq z\leq a+2L} K\left(\frac{x-z}{w}\right) dx = \int \sup_{-L\leq y\leq L} K\left(\frac{x-a-L-y}{w}\right) dx = \int \sup_{-L\leq y\leq L} K\left(\frac{v-y}{w}\right) dv,$$

(40) follows.

If K is nonincreasing on $(M_2, +\infty)$ and $x > M_2 w + L$, then $\sup_{-L\leq y\leq L} K\left(w^{-1}(x-y)\right) = K\left(w^{-1}(x-L)\right)$ and

$$\int_{M_2 w+L}^{\infty} \sup_{-L\leq y\leq L} K\left(\frac{x-y}{w}\right) dx = \int_{M_2 w+L}^{\infty} K\left(\frac{x-L}{w}\right) dx = w\int_{M_2}^{\infty} K(y)\, dy.$$

Similarily,

$$\int_{-\infty}^{M_1 w-L} \sup_{-L\leq y\leq L} K\left(\frac{x-y}{w}\right) dx = w\int_{-\infty}^{M_1} K(y)\, dy$$

and finally

$$\int \sup_{-L\leq y\leq L} K\left(\frac{x-y}{w}\right) dx \leq (M_2 w + L - M_1 w + L)\|K\|_\infty + w\int_{-\infty}^{M_1} K(x)\, dx + w\int_{M_2}^{\infty} K(x)\, dx,$$

which is (41). The unimodal case immediately follows from (40) and (41) with $M_1 = M_2$. □

Acknowledgments One author thanks Mathieu Sart for his helpful comments on an earlier version of the paper.

References

[1] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[2] S. Arlot and M. Lerasle. Why V = 5 is enough in V-fold cross-validation. *arXiv:1210.5830v2*, 2014.

[3] Y. Baraud. Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields*, 151:353–401, 2011.

[4] A. Berlinet and L. Devroye. A comparison of kernel density estimates. *Publications de l'Institut de Statistique de l'Université de Paris*, 38(3):3–59, 1994.

[5] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 65:181–237, 1983.

[6] L. Birgé. Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Ann. Inst. H. Poincaré Sect. B*, 20:201–223, 1984.

[7] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Institut Henri Poincaré, Probab. et Statist.*, 42:273–325, 2006.

[8] L. Birgé. Robust tests for Model Selection. *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner (M.Banerjee, F. Bunea, J. Huang, V. Koltchinskii and M. Mathuis,eds)*, IMS Collections – Volume 9:47–64, 2013.

[9] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97:113–150, 1993.

[10] L. Birgé and Y. Rozenholc. How many bins should be put in a regular histogram. *ESAIM Probab. Statist.*, 10:24–45, 2006.

[11] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation.* Springer-Verlag, New York, 2001.

[12] L.M. Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–55, 1973.

[13] L.M. Le Cam. On local and global properties in the theory of asymptotic normality of experiments. *In Stochastic Processes and Related Topics, Academic Press, New York.*, 1:13–54, 1975.

[14] N. Magalhães and Y. Rozenholc. An efficient algorithm for T-estimation. *http://hal.archives-ouvertes.fr/hal-00986229*, 2014.

[15] J.S. Marron. A Comparison of Cross-Validation Techniques in Density Estimation. *Ann. Statist.*, 15(1):152–162, 1987.

[16] P. Massart. *Concentration Inequalities and Model Selection.* Lecture on Probability Theory and Statistics. Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003 (J. Picard, ed.) Lecture Notes in Math. Springer, Berlin, 2007.

[17] T. Mildenberger and H. Weinert. The benchden package: Benchmark densities for non-parametric density estimation. *Journal of Statistical Software*, 46(14):1–14, 2012.

[18] M. Rudemo. Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics.*, 9(2):65–78, 1982.

[19] G. Walter and J. Blum. Probability density estimation using delta sequences. *Ann. Statist.*, 7:328–340, 1979.

[20] G.S. Watson and M.R. Leadbetter. Hazard analysis II. *Sankhya Ser. A.*, 26:101–116, 1965.

[21] P. Whittle. On the smoothing of probability density functions. *J. Roy. Statist. Soc. Ser. B.*, 20:334–343, 1958.

[22] B.B. Winter. Rate of strong consistency of two nonparametric density estimators. *Ann. Statist.*, 3(3):759–766, 1975.

A Supplementary material

We provide here additional simulations about the TVF based on the test statistic $T_{t,u}(\mathbf{X})$ designed by Baraud in [3] and given by (18). As in Section 4, we study the influence of V and we compare the TVF based on this test with classical VF procedures. The results are summarized in Table 3 and Figure 3 which are the analogues of Table 2 and Figure 2 respectively.

family	V	s_1	s_2	s_3	s_4	s_5	s_7	s_{12}	s_{13}	s_{22}	s_{23}	s_{24}
$\mathcal{F}_{\mathrm{R}}$	2	**2,89**	9,97	9,07	13,2	10,5	11	17,5	14,7	10,3	19,9	26,9
	5	4,33	9,68	8,61	12,4	9,87	10,4	17,1	**13,4**	9,37	17,8	24,7
	10	6,13	9,65	8,56	12,1	9,65	10,4	17	13,7	9,36	17,5	**24,3**
	20	9,28	**9,47**	**8,4**	**12**	**9,36**	**10,3**	**16,9**	14,2	**9,17**	**17,4**	24,6
	BR	2,20	9,94	9,27	12,98	10,53	11,14	17,85	14,63	10,37	17,98	25,15
$\mathcal{F}_{\mathrm{K}}$	2	15,6	29,4	5,69	5,07	**3,55**	4,24	27,2	20	3,97	10,3	18
	5	13,2	25,7	5,1	**4,94**	3,58	**3,97**	23	18,1	**3,85**	9,18	16,2
	10	12,9	24,8	5	5,02	3,86	4,01	22,2	17,7	3,87	9,04	**15,8**
	20	**12,7**	**24,4**	**4,98**	5,28	4,54	4,1	**21,6**	**17,6**	3,98	**8,98**	**15,8**
	UCV	15,86	22,20	5,57	6,16	3,74	4,10	18,80	17,16	3,88	9,52	15,91
$\mathcal{F}_{\mathrm{KR}}$	2	**2,87**	10	7,47	5,88	5,04	5,6	18,9	14,7	6,38	11,6	19,1
	5	3,68	**9,77**	6,81	**5,48**	**4,64**	**5,19**	17,7	**13,3**	**5,01**	9,3	16,4
	10	3,58	9,84	6,71	5,53	4,99	5,26	**17,6**	13,7	5,11	9,04	15,9
	20	3,79	9,84	**6,45**	5,65	5,31	5,83	**17,6**	14,6	5,22	**9,01**	**15,7**

Table 3: 1000 times Hellinger risks for the TVF procedure based on Baraud's test.

Influence of the test on the TVF

We compare here the performances of the best TVF procedure (among the five values of θ described above) derived from Birgé's test (9) against the one deduced from Baraud's test (18) (denoted $\widehat{s}_{\widehat{m}_{\mathrm{TVF}}}$). We show the conclusion of our study for the families $\mathcal{F}_{\mathrm{R}}$, $\mathcal{F}_{\mathrm{K}}$ and $\mathcal{F}_{\mathrm{KR}}$, $n = 500$ and $V = 2, 5, 10$ and 20. The results are very similar for other values of n. For the sake of clarity and to emphasize the similarity of both procedures in terms of Hellinger risk, we present for each family, for each V, the supremum and the infimum over $\mathcal{L}$ of the ratio

$$\Upsilon(s) = \left\{ \inf_{\theta\in\Theta} \overline{R}_n\left(\widehat{s}_{\widehat{m}(\theta)}, s, h^2\right) \Big/ \overline{R}_n\left(\widehat{s}_{\widehat{m}_{\mathrm{TVF}}}, s, h^2\right) \right\}.$$

If $\inf_{s\in\mathcal{L}} \Upsilon(s) \geq 1$ the TVF using Baraud's test behaves in a better way than the one using Birgé's test for all densities in $\mathcal{L}$ while if $\sup_{s\in\mathcal{L}} \Upsilon(s) \leq 1$ the opposite holds. The closer the two values, the more similar the quality of both procedures.

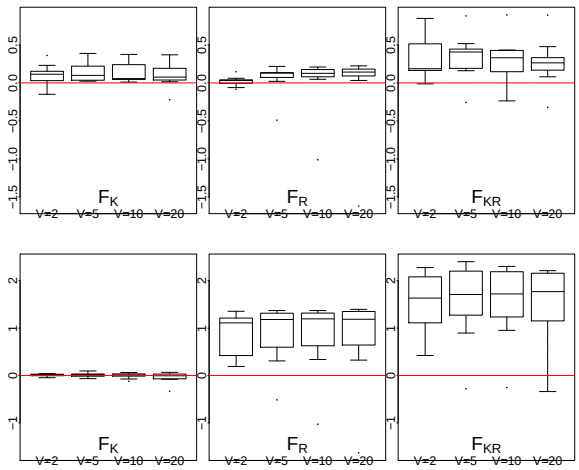

Figure 3: From left to right, the boxplots of $\overline{W}_s\left(\tilde{s}, \tilde{s}_{\mathrm{TVF}}, h^2\right)$ using families $\mathcal{F}_{\mathrm{K}}, \mathcal{F}_{\mathrm{R}}$ and $\mathcal{F}_{\mathrm{KR}}$ (up for $\tilde{s} = \widehat{s}_{\widehat{m}_{\mathrm{LSVF}}}$, down for $\tilde{s} = \widehat{s}_{\widehat{m}_{\mathrm{KLVF}}}$). Each subfigure shows the boxplot for $V = 2, 5, 10$ and 20. The horizontal red dotted line provides the reference value 0.

family	$\Upsilon(s)$	$V=2$	$V=5$	$V=10$	$V=20$
$\mathcal{F}_{\mathrm{R}}$	$\sup_s$	103,68	102,59	101,72	102,27
	$\inf_s$	98,16	100,07	99,59	99,13
$\mathcal{F}_{\mathrm{K}}$	$\sup_s$	102,78	100,80	100,92	105,10
	$\inf_s$	99,58	98,72	97,45	96,13
$\mathcal{F}_{\mathrm{KR}}$	$\sup_s$	116,71	115,80	116,79	116,73
	$\inf_s$	96,70	98,84	99,08	99,30

Table 4: Supremum and infimum of 100 times the ratio, see the text.

We see from this table that Baraud's and Birgé's test are very similar to process the TVF procedure for families $\mathcal{F}_{\mathrm{R}}$ and $\mathcal{F}_{\mathrm{K}}$. There is indeed no noticeable difference for these families, the largest gain (for a density in $\mathcal{L}$) being of 5% only. The procedure based on Baraud's test becomes much better for the family $\mathcal{F}_{\mathrm{KR}}$. We observe indeed that a potential gain of 15% appears (since the $\sup_s$ is close to 115%) while the loss is negligible (since the $\inf_s$ is close to 99%). Moreover, the ratios are quite similar when V increases. Finally, let us recall that the TVF procedure based on (9) is less time-consuming since it requires to compute only one integral instead of two for (18).